Adaptation de contexte basée sur la Qualité d'Expérience dans les réseaux Internet du Futur

Wael Cherif

ANNÉE 2013

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale MATISSE

présentée par

Wael CHERIF

Préparée à l'unité de recherche
INRIA - Rennes

Adaptation de contexte basée sur la Qualité d'Expérience dans les réseaux Internet du Futur

Thèse soutenue à Rennes le 19 juin 2013

devant le jury composé de :

André-Luc BEYLOT
Professeur, Université de Toulouse / *rapporteur*

Pascal LORENZ
Professeur, Université de Haute Alsace / *rapporteur*

César VIHO
Professeur, Université de Rennes 1 / *examinateur*

Daniel NEGRU
Maître de Conférences, Université de Bordeaux / *examinateur*

Mamadou SIDIBE
Ingénieur R&D, Viotech Communications / *examinateur*

Gerardo RUBINO
Directeur de Recherche, INRIA / *directeur de thèse*

Adlen KSENTINI
Maître de Conférences, Université de Rennes 1 / *co-directeur de thèse*

A ma famille…

Remerciements

Je souhaite rendre hommage et exprimer ma profonde gratitude à tous ceux qui, de près ou de loin, ont contribué à la réalisation et à l'aboutissement de cette thèse.

Les travaux présentés dans cette thèse ont fait l'objet d'une convention CIFRE entre la société Viotech Communications et le laboratoire INRIA de Rennes.

Je tiens surtout à remercier très sincèrement mes directeurs de thèse Messieurs Gerardo Rubino, Adlen Ksentini et Daniel Négru, dont l'expérience et la pédagogie ont été pour moi sources de motivation et de curiosité durant ces trois années.

Je tiens dans un premier temps à remercier Monsieur Gerardo Rubino, Directeur de recherche à l'INRIA, pour m'avoir confié ce travail de recherche, et de m'avoir donné la chance de travailler avec l'équipe Dionysos. Je remercie également Monsieur Adlen Ksentini, maitre de conférences à l'Université de Rennes 1, et co-encadrant de ce travail de thèse, pour ses idées et conseils, ainsi que pour son aide précieuse de tous les jours.

Je remercie également Monsieur Daniel Négru, maitre de conférences à l'Université de Bordeaux, ainsi que Monsieur Mamadou Sidibé, ingénieur R&D chez Viotech Communications, d'avoir codirigé cette thèse. Sans eux, cette thèse CIFRE n'aurait surement jamais vu le jour ; Je remercie aussi tous les collaborateurs du projet européen ALICANTE, spécialement Monsieur Michael Grafl, Docteur à l'Université de Klagenfurt, Monsieur Christian Timmerer, Professeur à l'Université de Klagenfurt et Daniel Renzi, assistant scientifique à l'Ecole Polytechnique Fédérale de Lausanne.

Je tiens également à témoigner toute ma reconnaissance aux Professeurs André-Luc Beylot de l'Université de Toulouse et Pascal Lorenz de l'Université de Haute Alsace pour l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de ce travail de recherche et membres du jury.

J'associe à ces remerciements, le président du jury, Monsieur César Viho, Professeur à l'Université de Rennes 1, qui m'a fait l'honneur de présider le jury de cette thèse.

Je tiens également à avoir une pensée sympathique pour toutes les personnes de l'équipe Dionysos du laboratoire de recherche INRIA de Rennes, qui savent si bien rendre agréable le cadre de travail et plus particulièrement à Anthony, Nanxing, Kamal, Mathieu, Fabienne, Kandaraj et Yassine.

Je termine par une profonde pensée à mes parents, ma sœur, mes proches et mes amis pour leur soutien sans faille durant ces trois ans.

Résumé

Pour avoir une idée sur la qualité du réseau, la majorité des acteurs concernés (opérateurs réseau, fournisseurs de service) se basent sur la Qualité de Service (*Quality of Service*). Cette mesure a montré des limites et beaucoup d'efforts ont été déployés pour mettre en place une nouvelle métrique qui reflète, de façon plus précise, la qualité du service offert. Cette mesure s'appelle la qualité d'expérience (*Quality of Experience*). La qualité d'expérience reflète la satisfaction de l'utilisateur par rapport au service qu'il utilise. Aujourd'hui, évaluer la qualité d'expérience est devenu primordiale pour les fournisseurs de services et les fournisseurs de contenus. Cette nécessité nous a poussés à innover et concevoir des nouvelles méthodes pour estimer la QoE. Dans cette thèse, nous travaillons sur l'estimation de la QoE (1) dans le cas des communications Voix sur IP et (2) dans le cas des services de diffusion Vidéo sur IP. Nous étudions les performances et la qualité des codecs iLBC, Speex et Silk pour la VoIP et les codecs MPEG-2 et H.264/SVC pour la vidéo sur IP. Nous étudions l'impact que peut avoir la majorité des paramètres réseaux, des paramètres sources (au niveau du codage) et destinations (au niveau du décodage) sur la qualité finale. Afin de mettre en place des outils précis d'estimation de la QoE en temps réel, nous nous basons sur la méthodologie *Pseudo-Subjective Quality Assessment*. La méthodologie PSQA est basée sur un modèle mathématique appelé les réseaux de neurones artificiels. En plus des réseaux de neurones, nous utilisons la régression polynomiale pour l'estimation de la QoE dans le cas de la VoIP.

Abstract

Quality of Experience (QoE) is the key criteria for evaluating the Media Services. Unlike objective Quality of Service (QoS) metrics, QoE is more accurate to reflect the user experience. The Future of Internet is definitely going to be Media oriented. Towards this, there is a profound need for an efficient measure of the Quality of Experience (QoE). QoE will become the prominent metric to consider when deploying Networked Media services. In this thesis, we provide several methods to estimate the QoE of different media services: Voice and Video over IP. We study the performance and the quality of several VoIP codecs like iLBC, Speex and Silk. Based on this study, we proposed two methods to estimate the QoE in real-time context, without any need of information of the original voice sequence. The first method is based on polynomial regression, and the second one is based on an hybrid methodology (objective and subjective) called *Pseudo-Subjective Quality Assessment*. PSQA is based on the artificial neural network mathematical model. As for the VoIP, we propose also a tool to estimate video quality encoded with MPEG-2 and with H.264/SVC. In addition, we study the impact of several network parameters on the quality, and the impact of some encoding parameters on the SVC video quality. We also made performance tests of several SVC encoders and proposed some SVC encoding recommendations.

Table des matières

Liste des figures

Liste des tableaux

Acronymes

Acronyme	Signification
ACELP	Algebraic Code Excited Linear Prediction
ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
AMR	Adaptive Multi Rate
ANN	Artificial Neural Network
ASO	Arbitrary Slice Ordering
AVC	Advanced Video Coding
BI-LPC	Block-Independent Linear Predictive Coding
BL	Base Layer
CABAC	Context Adaptive Binary Arithmetic Coding
CAVLC	Context Adaptive Variable Length Coding
CBD	City Block Distance
CBR	Constant BitRate
CELP	Code-Excited Linear Prediction
CGS	Coarse-Grain Scalability
CNG	Comfort Noise Generator
CS-ACELP	Conjugate Structure Algebraic Code Excited Linear Prediction
DCR	Degradation Category Rating
DCT	Discrete Cosine Transformation
DMOS	Degradation Mean Opinion Score
DMOS	Differential Mean Opinion Score
DP	Data Partitioning
DPCM	Differential Pulse Code Modulation
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DTX	Discontinuous Transmission
DVB	Digital Video Broadcasting
EL	Enhancement Layer
ETSI	European Telecommunications Standards Institute
EVD	Energy Variation Descriptor
FEC	Forward Error Correction
FMO	Flexible Macroblock Ordering
FPS	Frames Per Second
FR	Full-Reference
GGD	Generalized Gaussian Density
GMM	Gaussian Mixture Models

GOP	Group Of Picture
GSM	Global System for Mobile
GSM-EFR	GSM Enhanced Full Rate
GSM-FR	GSM Full Rate
GSM-HR	GSM Half Rate
HD	High Definition
HLS	HTTP Live Streaming
HMM	Hidden Markov Models
HTTP	HyperText Transfer Protocol
HVS	Human Visual System
IEC	International Electrotechnical Commission
iLBC	Internet Low Bitrate Codec
INMD	In-service Non-intrusive Measurement Devices
IP	Internet Protocol
IPTV	IP Television
ISO	International Standards Organization
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
JPEG	Joint Photographic Experts Group
JSVM	Joint Scalable Video Model
LAN	Local Area Network
LAR	Log Area Ratio
LBS	Loss Burst Size
LD-CELP	Low Delay Code Excited Linear Prediction
LHS	Local Harmonic Strength
LLR	Log-Likelihood Ratio
LPC	Linear Predictive Coding
LR	Loss Rate
MANE	Media Aware Network Element
MGS	Medium-Grain Scalability
MIC	Modulation d'Impulsions Codées
MICDA	Modulation par impulsion et codage différentiel adaptatif
MLBS	Mean Loss Burst Size
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MP-MLQ	Multipulse Maximum Likelihood Quantization
MSE	Mean Square Error
MSS	Microsoft Smooth Streaming
NAL	Network Abstraction Layer
NALU	Network Abstraction Layer Unit
NR	No-Reference

NTIA	National Telecommunications and Information Administration
PAMS	Perceptual Analysis Measurement System
PAQM	Perceptual Audio Quality Measure
PC	Pair Comparison
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PLC	Packet Loss Concealment
POLQA	Perceptual Objective Listening Quality Analysis
PSNR	Peak Signal-to-Noise Ratio
PSQA	Pseudo-Subjective Quality Assessment
PSQM	Perceptual Speech Quality Measure
QCIF	Quarter of Common Intermediate Format
QoE	Quality of Experience
QoS	Quality of Service
QP	Quantization Parameter
RD	Rate-Distortion
RELP	Residual Excited Linear Prediction
RR	Reduced-Reference
SAD	Speech Activity Detection
SAMVIQ	Subjective Assessment Methodology for VIdeo Quality
SCACJ	Stimulus Comparison Adjectival Categorical Judgement
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SEI	Supplementary Enhancement Information
SI	Spatial Information
SNR	Signal-to-Noise-Ratio
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity and Image Quality
SSM	Single Stimulus Method
SSNR	Segmental Signal-to-Noise-Ratio
SVC	Scalable Video Coding
TCP	Transmission Control Protocol
TI	Temporal Information
UDP	User Datagram Protocol
VAD	Voice Activity Detection
VBR	Variable Bit Rate
VCL	Video Coding Layer
VOD	Video On Demand
VoIP	Voice over IP
VQM	Video Quality Metric
WB-PESQ	Wide Band - Perceptual Evaluation of Speech Quality

Chapitre 1

1 Introduction

1.1 Motivations

Au fil des années, les applications multimédias ont conquis plusieurs segments du domaine des télécommunications. Aujourd'hui, nous avons affaire à des services multimédias dans de nombreux domaines, à commencer par les différents systèmes de télévision numérique (par exemple DVB), la vidéo-téléphonie, la vidéo à la demande (VOD), des services de télévision sur protocole Internet (IPTV), la voix sur IP (VoIP) ou tout simplement, les services de partage vidéo comme YouTube[1] ou Dailymotion[2]. Le développement de ces services et l'optimisation de bout en bout de ces systèmes sont étroitement liés à la perception de la qualité par l'utilisateur et à sa satisfaction à l'égard du service rendu. Dans ce sens, il y a un profond besoin d'une mesure qui permette de refléter la satisfaction et la perception des utilisateurs. En effet, les fournisseurs des services médias sont de plus en plus intéressés par l'évaluation de la performance de leurs services fournis telle que perçue par les utilisateurs finaux, afin d'améliorer et de mieux comprendre les besoins de leurs clients. Les opérateurs réseau sont aussi intéressés par cette mesure afin d'optimiser les ressources du réseau et même (re)configurer les paramètres réseaux pour accroître la satisfaction des utilisateurs.

Il existe plusieurs façons pour obtenir des informations sur la qualité perçue. D'une part, il y a des évaluations subjectives effectuées dans des laboratoires parfaitement équipés pour enquêter sur la perception de l'utilisateur final. D'autre part, il y a des mesures objectives de la qualité, qui sont souvent utilisées pour étudier les paramètres mesurables de l'ensemble du système, décrivant d'une façon technique la qualité de service (QoS). Cependant, ces paramètres ne peuvent pas décrire toutes les variables qui influencent la perception de la qualité du côté de l'utilisateur final. Pour cette raison, une nouvelle mesure, appelée la « qualité d'expérience (QoE) », a été définie afin de refléter la qualité perçue par les utilisateurs finaux.

La définition de la QoE est étroitement liée à la perception subjective de l'utilisateur final. La QoE est décrite par l'ITU-T comme étant « l'acceptabilité globale d'une application ou d'un service, tel qu'il est perçu subjectivement par l'utilisateur final », qui « peut être influencée par le contexte et les attentes des utilisateurs » [1].

[1] www.youtube.fr

[2] www.dailymotion.fr

Les recherches sur la Qualité d'Expérience (QoE) sont souvent basées sur des études subjectives. Dans ces études subjectives, les utilisateurs notent la qualité perçue d'un service ou d'une application. En règle générale, ces études sont effectuées dans un laboratoire spécialisé. Cependant, ces tests subjectifs sont fastidieux et coûteux. De plus, ce genre de test ne peut s'appliquer dans un système à temps-réel, tels que la plupart des services médias.

Dans ce contexte, les recherches se sont concentrées sur des nouvelles méthodes qui essaient d'approximer et d'estimer la qualité d'expérience, d'une façon objective et qui peuvent être utilisées dans des contextes temps-réel. Le principal inconvénient des solutions existantes réside dans le fait qu'elles ne sont pas en corrélation avec les tests subjectifs, et par conséquent, elles ne peuvent correctement refléter la perception de l'utilisateur final.

L'objectif principal de cette thèse est de proposer des méthodes performantes qui permettent d'estimer, d'une façon précise et en temps-réel, la qualité perçue par des utilisateurs d'un service donné. Ainsi, nous nous sommes concentrés sur deux services multimédia très populaires : la diffusion de Vidéo sur IP et la Voix sur IP. Nous avons traité dans nos recherches plusieurs types de codecs (les plus fréquemment utilisés), tels que le MPEG-2 et le H.264 pour le service de vidéo sur IP ; et Speex, iLBC et Silk pour celui de VoIP.

1.2 Contributions

Dans le cadre de cette thèse, nous allons nous baser sur la définition de la qualité d'expérience (fournie par l'ITU) pour proposer des solutions qui permettent son estimation. Nos contributions concernent les deux services médias les plus utilisés : la Vidéo sur IP et la Voix sur IP ; et consistent à utiliser la méthodologie PSQA, basée sur les réseaux de neurones, pour estimer, d'une façon objective et en temps réel, la qualité perçue par l'utilisateur final du service proposé.

Nos contributions peuvent être résumées en quatre parties :

- Une comparaison des performances des méthodes objectives existantes qui permettent l'évaluation de la qualité des vidéos encodées en MPEG-2. Nous avons mis en place une plateforme de test pour faire du streaming vidéo, et nous avons généré des pertes réseaux pour détériorer la qualité de la vidéo. La qualité des vidéos reçues par le client (utilisateur final) a été évaluée subjectivement et objectivement. Nous avons comparé ensuite les résultats de l'évaluation subjective avec les estimations fournies par PSQA et d'autres méthodes objectives [1].

- Une étude des performances du codec H.264/SVC qui consiste à analyser les performances de ce codec, et ce en étudiant l'impact que peut avoir les différents paramètres du codec (tels que le facteur de quantification, le nombre de couches, etc.) sur le débit binaire et sur la qualité finale de la vidéo. Nous avons également comparé les performances de différents encodeurs SVC [4] [5].

- Une proposition d'un outil pour estimer la QoE des vidéos SVC diffusées. Un test subjectif a été réalisé pour vérifier l'impact de plusieurs paramètres réseaux et certains paramètres liés

à l'encodeur sur la qualité perçue. A partir de ces données, nous avons proposé un outil, basé sur les réseaux de neurones, pour estimer la QoE [3].

- Une étude des performances des encodeurs de la voix sur IP les plus récents et les plus utilisés tels que iLBC, Speex et Silk. Nous nous sommes basés sur un outil objectif intrusif standardisé, appelé PESQ, pour l'estimation de la qualité. A partir de l'estimation de qualité des différents encodeurs, sous différentes conditions réseaux, nous avons proposé deux méthodes efficaces et précises pour estimer la qualité perçue lors d'une communication VoIP :
  - Une solution basée sur la régression polynomiale : la solution est un ensemble de fonctions polynomiales, qui prend comme paramètres : les caractéristiques du codec et les conditions du réseau ; et qui fournit comme résultat une estimation de la QoE.
  - Une solution basée sur les réseaux de neurones : la solution est un ensemble de réseaux de neurones, qui prennent les conditions du réseau et les caractéristiques du codec en entrée et fournissent en sortie une estimation de la QoE [2].
- Une intégration des solutions proposées dans un module complet de Monitoring au niveau des terminaux utilisateurs, dans le cadre du projet européen ALICANTE.

1.3 Organisation de la thèse

Cette thèse est organisée comme suit :

Le chapitre 2 est un état de l'art sur la Vidéo sur IP. Nous expliquons dans ce chapitre le principe de codage vidéo et la structure des séquences vidéo codées. Nous nous sommes intéressés aux codecs vidéo les plus utilisés qui sont le MPEG-2 et le H.264. Dans la seconde partie de ce chapitre, nous discuterons des différents paramètres qui peuvent avoir un impact sur la qualité des séquences vidéo et nous présentons quelques outils et méthodes pour estimer la qualité.

Le chapitre 3 concerne les principes de base de la communication par Voix sur IP. Nous présentons dans ce chapitre le fonctionnement des codecs VoIP : de l'acquisition à la reconstruction du signal « parole ». Dans ce chapitre, nous présentons les différents codecs les plus utilisés pour la VoIP et l'impact que peuvent avoir certains paramètres réseaux sur la qualité perçue par les utilisateurs. Nous citons ensuite des méthodes subjectives et objectives pour la mesure de la qualité des communications VoIP.

Le chapitre 4 est réservé à la méthodologie « *Pseudo-Subjective Quality Assessment* ». Dans ce chapitre, nous expliquons le principe des réseaux de neurones artificiel et son utilisation dans le cas d'estimation de la « *Quality of Experience* ». Nous évaluons dans ce chapitre les performances de la méthodologie PSQA en la comparant avec d'autres méthodes d'évaluations objectives et subjectives des vidéos MPEG-2.

Le chapitre 5 présente nos contributions sur l'étude des performances du codec H.264/SVC. Dans ce chapitre, nous présentons des recommandations d'encodage SVC et nous comparons différents encodeurs SVC.

Le chapitre 6 inclut une étude de l'impact que peuvent avoir certains paramètres SVC tels que l'encodage CGS/MGS, le nombre de couche, le débit binaire, … Nous proposons dans ce chapitre des méthodes d'évaluation, basées sur la méthodologie PSQA, de la qualité des flux vidéo SVC.

Le chapitre 7 détaille les contributions sur l'estimation de la qualité des communications VoIP. Nous présentons en premier lieu les expérimentations mises en place pour étudier l'impact que peuvent avoir certains paramètres réseaux sur la qualité. A partir des résultats d'évaluation de la qualité, nous présentons deux méthodes objectives qui permettent d'estimer la QoE en temps réel : la première méthode est basée sur la régression polynomiale ; la seconde méthode est basée sur les réseaux de neurones. Nous comparons ensuite les résultats obtenus des différentes méthodes.

Le chapitre 8 montre l'utilité de nos contributions dans le cadre du projet européen ICT-ALICANTE. Nous présentons, dans ce chapitre, le module à mettre en place au niveau de l'environnement utilisateur pour estimer la qualité d'expérience en utilisant nos méthodes proposées, et en se basant sur la mesure des métriques nécessaires.

Finalement, le chapitre 9 conclut cette thèse, et présente les différentes perspectives et directions des futurs travaux de recherche.

Chapitre 2

2 La vidéo sur IP

En règle générale, la numérisation d'une source vidéo génère une grande quantité de données. L'utilisation de données sous leurs formes numériques ne serait pas possible aujourd'hui sans la compression préalable de celles-ci. Et ceci pour plusieurs raisons :

- Les capacités de stockages des utilisateurs, même si elles ne cessent d'augmenter, ne sont pas infinies.
- La durée des transmissions de ces données numériques est conditionnée par le débit du réseau qui est utilisé et qui est parfois relativement faible.

Afin de faire face à la transmission et au stockage de la vidéo numérique, plusieurs techniques de compression vidéo ont été développées. Ces techniques sont communément appelées codecs vidéo. Le terme vient du processus de codage (ou compression) et de décodage (décompression) d'un signal vidéo pour le stockage et/ou la transmission. Les codecs visent à maximiser l'efficacité de compression et la qualité d'image, tout en minimisant la complexité de calcul. Ils permettent une compression en supprimant les informations redondantes à partir du signal d'origine. Il existe deux principaux types de redondance en vidéo :

- **Redondance d'image :** Les pixels d'une image ne sont pas complètement indépendants, ils ont une étroite corrélation avec leurs voisins, à la fois dans le même cadre (redondance spatiale) ou à travers des images successives (redondance temporelle). Dans une certaine mesure, un pixel peut être prédit à partir de ses voisins.
- **Redondance psycho-visuel :** Cette redondance est en rapport avec le système visuel humain. Tout comme l'oreille a une réponse en fréquence limitée, le couple œil-cerveau a des résolutions limitées. Il y a une limite (spatiale) aux détails que l'œil peut percevoir. Nous disons alors que l'œil est plus sensible aux basses fréquences (zones plates) qu'aux hautes fréquences (des zones de texture). Temporellement, il y a aussi une limite à la capacité de l'œil-cerveau pour suivre les images en mouvement rapide. Enfin, il y a des limites sur le nombre de couleurs que les êtres humains sont capables de distinguer.

Dans ce chapitre, nous expliquons la base du codage vidéo. Nous montrons comment une séquence vidéo est structurée et quels sont les différents types d'images qui composent une séquence vidéo. Nous présentons après les codecs vidéo les plus utilisés pour la IPTV et la TV-HD.

2.1 Principes des codecs vidéo

Les codecs vidéo standards sont composés de deux procédés principaux : la transformation spatiales de blocs et la compensation de mouvement. Les sections suivantes présentent une vue d'ensemble de ces processus, leurs objectifs et certaines techniques spécifiques.

2.1.1 Transformation spatiale

Le processus de transformation spatiale peut être illustré par le codage d'une image fixe exercé en vertu de la norme JPEG [2]. Tout d'abord, l'image est divisée en blocs de 8x8 pixels. Ensuite, une transformation dite DCT (*Discrete Cosine Transformation*) est appliquée sur le bloc. Le processus DCT est un processus réversible qui fait correspondre la représentation normale 2-D de l'image avec sa représentation dans le domaine fréquentiel. Le résultat de la DCT est également un bloc de coefficients de taille 8x8.

Chaque coefficient représente la contribution d'une fonction de base DCT pour le bloc d'image d'origine. DCT tend à concentrer l'énergie dans quelques coefficients (principalement des fréquences basses) et bien d'autres se trouvent à proximité ou égal à zéro. Ainsi, la compression initiale est réalisée par filtration sur ces coefficients proches de zéro. Les coefficients restants dans le bloc DCT sont ensuite quantifiés avec des résolutions différentes, en fonction de l'importance des fréquences pour la perception visuelle humaine. Cela signifie que nous utilisons plus de bits pour quantifier les coefficients de basses fréquences que ceux des hautes fréquences. Les coefficients quantifiés sont balayés en zigzag et traités en utilisant des techniques d'encodage entropique sans perte, connues sous le nom de codage à longueur variable et le codage de *Huffman*.

L'étape de quantification est une étape importante pour le contrôle de qualité des images. Cette étape pilote le compromis entre la qualité et le débit de l'image. En effet, plus le paramètre de quantification (*Quantization Parameter* QP) est élevé, plus il y a de coefficients nuls engendrés par la division entière, et plus important sera le rendement du codage de Huffman. Il est clair qu'un taux de compression élevé engendre une dégradation importante de la qualité visuelle de l'image restituée. L'algorithme de contrôle de débit est chargé de piloter le paramètre de quantification pour contrôler le débit généré par le codeur tout en minimisant la dégradation de la qualité visuelle.

2.1.2 Codage différentiel et la compensation de mouvement

Le principe clé de la compression vidéo numérique consiste à éliminer la redondance autant que possible dans le contexte d'utilisation. Les principales techniques qui permettent de diminuer la redondance temporelle d'une séquence vidéo sont : le codage différentiel et la compensation de mouvement. Ces techniques sont utilisées par les différents algorithmes MPEG.

- Le codage temporel différentiel est un moyen de codage assez naturel. Il part du constat que deux images successives I_t *et* I_{t+1} sont fortement semblables, et qu'il est plus avantageux de coder I_t puis $(I_{t+1} - I_t)$ que les images I_t *et* I_{t+1} de manière indépendante.

- La technique de compensation de mouvement part du constat qu'une séquence est souvent composée d'objets et de personnages qui traversent le champ de la caméra (*travelling*). De tels évènements peuvent être facilement modélisés par la translation d'une portion de l'image I_t

de la séquence vers une autre portion d'une image I_{t+1} de la séquence. Dans ce cas, pour avoir un gain de quantité de données à envoyer, il suffira de transmettre le vecteur translation plutôt que la totalité du bloc (ou de la différence entre les deux blocs dans le cas d'un codage différentiel). Il existe de nombreuses techniques pour estimer le mouvement, la technique utilisée par les normes MPEG est celle d'appariement de blocs, appelée *Block Matching*.

2.2 Structure d'une séquence vidéo codée

2.2.1 Hiérarchie des images

La plupart des codecs standardisés sont basés sur la prédiction par compensation de mouvement et le codage différentiel. Les images codées de la séquence sont décomposées en trois types différents :

- Les images de type « *intra* », notées *I*, qui seront codées sans prédiction et sans compensation de mouvement par un algorithme de compression.
- Les images de type « *prédites* », notées *P*, qui seront prédites à partir des images *I* ou *P* précédentes en utilisant une compensation de mouvement (*Forward prediction*). Pour améliorer la qualité de reconstruction, la différence entre l'image originale et l'image prédite sera également codée. Le codage utilisé pour coder cette différence sera le même que celui utilisé pour coder les images *I*.
- Les images de type « *bi-directionelles* », notées *B,* qui seront également prédites avec deux compensations de mouvement, l'une provenant d'une image *I* ou *P* passée, l'autre provenant d'une image *I* ou *P* future (*Bidirectional predection*). Comme pour les images *P*, la différence entre l'image prédite et l'image originale de la séquence sera également codée. Les images de type *B* sont utilisées pour fournir des images de qualité supérieure aux images de type *P*, par contre le codage des images *B* est plus coûteux car le nombre de vecteurs mouvements est double.

La succession des images de type *I*, *P* et *B* forme un groupe d'images (appelé GOP pour *Group Of Picture*). La succession des images *I*, *P* et *B* est au choix du codeur. La séquence codée obtenue est appelée *bitstream*. Il est à noter qu'un *bitstream* est un flux de bits, constitué par un ensemble d'éléments organisés en une hiérarchie et un ordre défini par le codeur qui l'a généré. La Figure 2.1 illustre la structure d'un GOP ainsi que les prédictions des images *P* et *B*.

L'un des inconvénients de cette hiérarchie entre les images, c'est que le *bitstream* est très sensible aux pertes. Habituellement, une erreur dans une image *I* se propage jusqu'à la prochaine image *I* (à cause de l'interdépendance des autres images du GOP). Les erreurs dans les trames *P* se propagent jusqu'à la prochaine trame *I* ou *P*. Les erreurs dans les trames *B* ne se propagent pas. Plus il y a des trames *I* dans la vidéo, plus la vidéo est robuste aux erreurs. Cependant, avoir un grand nombre de trames *I* augmente la taille de la vidéo. Afin d'économiser la bande passante, les vidéos destinées à la diffusion sur Internet ont souvent de grandes tailles de GOP. En MPEG-2, la taille du GOP est d'environ 20 images, alors que dans le format MPEG-4, il peut s'élever jusqu'à 250 images.

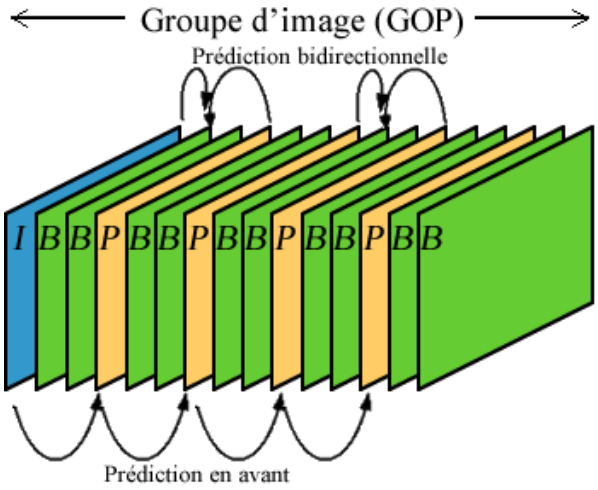


Figure 2.1. Exemple d'un groupe d'image GOP

2.2.2 Codage en couches

Un autre concept important dans le codage vidéo est le concept de codage en couches. Le codage vidéo en couches (ou *scalable*) est une technique dans laquelle la séquence d'origine est comprimée en un ensemble de sous-flux (*substreams*). L'idée est d'avoir une couche de base avec les informations essentielles (et de moins bonne qualité) et une ou plusieurs couches d'amélioration, ajoutées à la couche de base. Ces couches d'amélioration permettent d'augmenter le nombre d'images par seconde (*Frames Per Second* FPS) et/ou la qualité.

Le codage en couches est très utile dans le contexte d'un réseau avec une protection d'erreur inégale et lors de la multidiffusion (multicast) pour des clients hétérogènes. Dans un réseau avec protection d'erreur inégale, la couche de base peut être envoyée avec une priorité plus élevée que la (les) couche(s) d'amélioration(s) afin de la protéger contre les pertes en cas de congestion. Dans le cas du streaming multicast, chaque client peut avoir une vitesse d'accès différente. Avec le codage scalable, chaque client demande le maximum de couches qu'il peut effectivement recevoir. Si un client a plus de bande passante qu'un autre, alors il peut demander plus de couches et avoir des vidéos avec un taux plus élevé et une meilleure qualité.

Dans la suite, nous allons décrire quelques principaux codecs vidéo disponibles pour la diffusion de vidéo sur les réseaux IP.

2.3 Aperçu sur les standards MPEG

MPEG, acronyme de *Moving Picture Experts Group*, est un groupe de travail de l'organisation internationale de normalisation (*International Standards Organization ISO*) et de la Commission électrotechnique internationale (*International Electrotechnical Commission IEC*). L'objectif principal de ce groupe est la définition de normes pour la compression numérique des signaux vidéo et audio. MPEG a produit plusieurs normes depuis sa création en 1988 telles que : MPEG-1, MPEG-2 et MPEG-4.

Les algorithmes MPEG compressent les données pour former de petits morceaux qui peuvent être facilement transmis et puis décompressés. Les codecs MPEG atteignent des taux de compression élevés,

en stockant uniquement les changements d'une image à l'autre, au lieu de chaque trame. Les informations vidéo sont ensuite codées en utilisant la technique DCT (transformée en cosinus discrète). Les codecs MPEG utilisent une compression avec perte, car certaines données sont supprimées, mais la suppression de ces données est généralement imperceptible par l'œil humain.

Le flux MPEG contient un certain nombre de structures et d'éléments comportant chacun leur propre label comme illustré sur la Figure 2.2. Une connaissance minimale de la terminologie nous sera utile pour une meilleure compréhension des différents composants des flux vidéo MPEG. Le flux consiste en une ou plusieurs pistes audio et un flux vidéo. Le flux vidéo est constitué de GOP. Chaque GOP est associé à un *timecode* qui permet au décodeur d'assurer la synchronisation. Chaque image (*frame*) du GOP est encodée d'une des quatre façons possibles. Chacune de ces frames une fois encodée est divisée en tranches/bandes (*slice*) qui sont divisées à leur tour en macro-blocks. Un macro-block est un groupe d'éléments de base appelé blocs, et le bloc lui-même est un groupe de 8 x 8 pixels. Ce groupe peut contenir soit l'information chroma (CrCb) soit luma (Y).

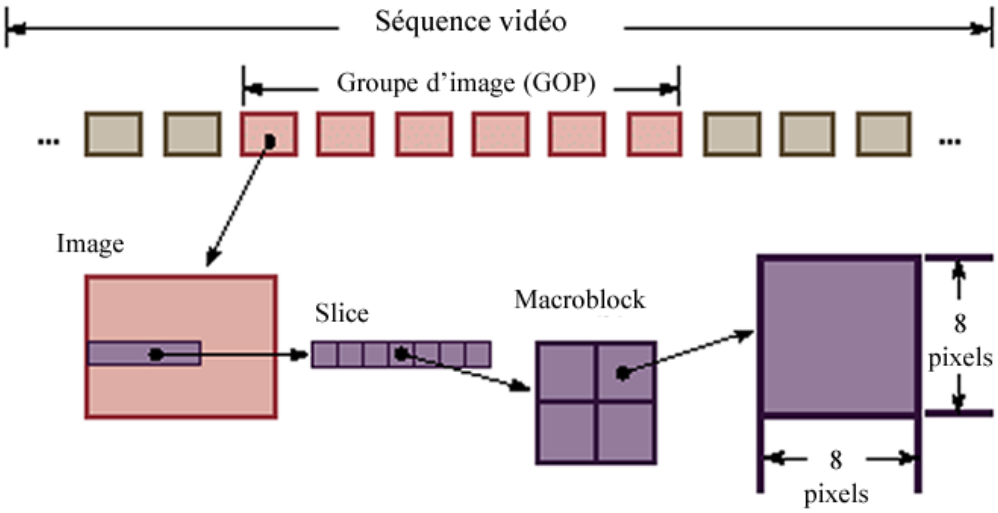


Figure 2.2. Syntaxe MPEG

2.3.1 MPEG-1/2

MPEG-1 [3] a été publié en 1993 et est conçu pour le stockage de médias numériques. MPEG-1 encode les vidéos non-entrelacées à des débits pouvant atteindre 1,5 Mbit/s. En 1994, une deuxième norme, appelée MPEG-2, a été définie [4]. MPEG-2 vise des applications à haute définition et des débits allant jusqu'à 30 Mbit/s. MPEG-2 est largement utilisé pour les DVD et la télévision numérique.

MPEG-1 et MPEG-2 sont conçus pour différents signaux d'entrée, différents taux de sortie et différents formats [5]. Ils fournissent également des capacités différentes pour le multiplexage audio et vidéo. Néanmoins, ils reposent tous deux sur les mêmes principes et techniques de codage.

MPEG-2 peut être considéré comme une version améliorée, de la norme MPEG-1, en termes de qualité. MPEG-2 peut encoder l'audio/vidéo dans des résolutions plus élevées et utiliser des débits plus élevés, comparés à la norme MPEG-1.

2.3.2 MPEG-4

MPEG-2 a eu beaucoup de succès et est la base de beaucoup d'applications actuellement répandues. Le groupe MPEG a commencé à travailler sur une nouvelle norme de compression vidéo. Le résultat de cet effort est connu comme MPEG-4, une norme partagée en plusieurs parties. Les différentes parties de la norme MPEG-4 couvrent des aspects comme le codage vidéo, codage audio, les problèmes liés aux systèmes de transport, réseau, etc.

2.3.3 H.264 ou MPEG-4/AVC

MPEG-4 a plusieurs profils comme cité ci-dessus. H.264 [6] fait référence à la partie 10 du profil MPEG-4. Son principal objectif est une compression efficace et robuste des images vidéo pour des applications telles que le stockage, la visioconférence, la visiophonie, la diffusion et le streaming sur une grande variété de technologies de transport.

H.264 (MPEG-4 Part 10) comprend de nouvelles techniques qui lui permettent de compresser beaucoup plus efficacement les vidéos que les normes précédentes (H.261, MPEG-1, MPEG-2) et fournit plus de flexibilité aux applications dans un grand nombre d'environnements réseau. Le standard de codage vidéo H.264 vise à gagner jusqu'à 50% de la bande passante utilisée par MPEG-2 pour une qualité visuelle équivalente.

La performance supérieure de H.264 vient de l'effet d'une série d'innovations algorithmiques. Pour la prédiction de mouvement, ces améliorations comprennent l'utilisation des tailles de bloc petites et variables, avec un quart de pixel de précision, une prédiction spatiale sur le bord des blocs voisins pour un codage « intra » et l'utilisation de plusieurs images de référence [7]. D'autres améliorations sont notamment la transformation effectuée sur des blocs de petites tailles et l'utilisation des méthodes de codage entropique avancées tels que CABAC et CAVLC [7].

En plus d'une amélioration de l'efficacité de codage, la norme H.264 a été spécialement conçue pour le transport de vidéo sur une variété de technologies de transport. La conception de la norme H.264 répond à ce besoin de flexibilité et de personnalisation en séparant l'information codée en deux couches : la couche de codage vidéo (*Video Coding Layer* VCL) et la couche d'abstraction de réseau (*Network Adstraction Layer* NAL) [7], comme indiqué dans la Figure 2.3. VCL représente l'information vidéo codée de la manière la plus efficace possible. La couche NAL formate les données VCL et les organise dans des éléments avec les en-têtes appropriés pour le transport ou le stockage sur une grande variété de technologies.

En ce qui concerne la robustesse de la norme contre les erreurs, H.264 inclut plusieurs fonctionnalités avancées. Tout d'abord, les éléments de syntaxe clés de la structure H.264 sont des tranches de taille flexible. Chaque tranche est transportée dans un unique paquet (appelé *NAL Unit* ou NALU), qui est ensuite passée à la couche transport pour être encapsulé (par exemple en *Real-Time Transport Protocol* RTP). Cette flexibilité permet d'adapter efficacement le flux codé à la technologie de

transport utilisée. H.264 comprend également les concepts de partitionnement de données, l'ordonnancement flexible des macro-blocs et l'ordonnancement arbitraire des tranches [8].

Le partitionnement des données (*Data Partitioning* DP) donne la possibilité de séparer les éléments de syntaxe d'importance plus ou moins élevée dans différents paquets de données. Ceci permet d'appliquer un niveau de protection inégal aux erreurs en fonction de l'importance des données et d'améliorer ainsi la fiabilité du flux. L'ordonnancement flexible des macro-blocs (*Flexible Macroblock Ordering* FMO) et l'ordonnancement arbitraire des tranches (*Arbitrary Slice Ordering* ASO) sont des techniques de restructuration de l'ordonnancement des régions fondamentales de l'image (macroblocs). Ces techniques sont généralement utilisées pour améliorer la résistance aux erreurs et aux pertes.

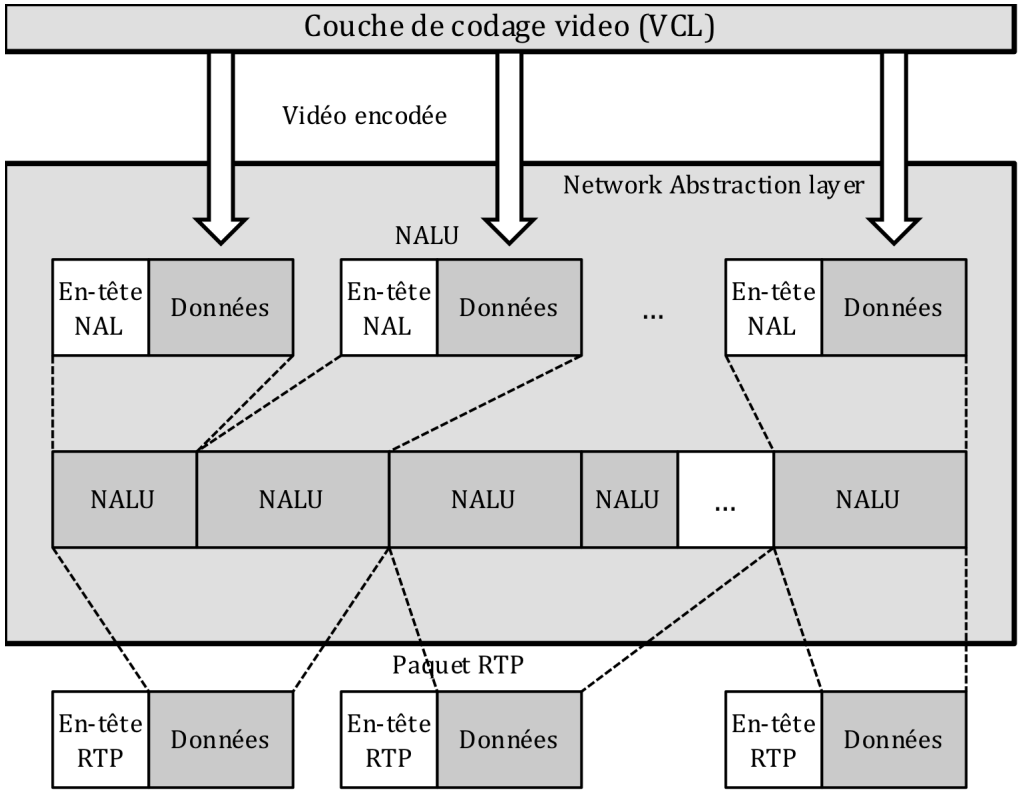


Figure 2.3. Structure des couches VCL et NAL

2.3.4 H.264 Scalable Video Coding

L'amendement *Scalable Video Coding* (SVC) [9], de la norme *MPEG-4 Advanced Video Coding* (AVC) suit un schéma de codage de couche comprenant une couche de base et une ou plusieurs couches d'amélioration avec des dimensions différentes. La Figure 2.4 illustre le concept de codage en couches en montrant un émetteur codant une séquence vidéo en trois couches complémentaires. Par conséquent, les récepteurs peuvent sélectionner et décoder un nombre différent de couches - chacune correspondant à des caractéristiques vidéo distinctes - en conformité avec les conditions et les contraintes du réseau et l'appareil lui-même.

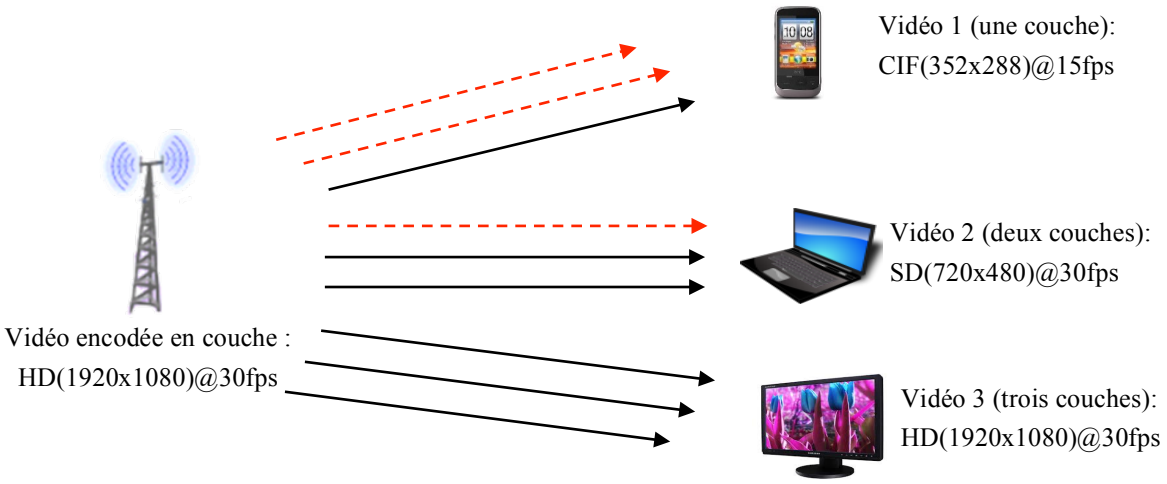


Figure 2.4. L'adaptation des vidéos codées en couche

Trois modes de codage scalable sont pris en charge et peuvent être combinés en un seul flux codé, à savoir la scalabilité spatiale, la scalabilité temporelle et la scalabilité en qualité (*Signal-to-Noise Ratio* SNR) :

- Scalabilité spatiale (taille de l'image) : la vidéo est codée sur plusieurs résolutions spatiales. Les données et les échantillons décodés de résolution minimale peuvent être utilisées pour prédire les données ou les échantillons de résolution plus élevée afin de réduire le débit binaire pour coder les résolutions supérieures.

- Scalabilité temporelle (taux d'image / *frame rate*) : les dépendances de compensation de mouvement sont structurées de telle sorte que des images complètes (i.e. leurs paquets associés) peuvent être écartées à partir du flux binaire. Notons que la scalabilité temporelle est déjà utilisée par H.264/MPEG-4 AVC. SVC n'a fourni que des informations d'amélioration supplémentaire, sous forme de « *Prefix NAL Units* », en-têtes des NALU des couches d'amélioration ou des messages *Supplemental Enhancement Information* SEI, pour améliorer son utilisation.

- Scalabilité SNR / Qualité / Fidélité : la vidéo est codée avec une résolution spatiale unique mais à des qualités différentes. Les données et les échantillons décodés de qualité inférieure peuvent être utilisés pour prédire les données ou les échantillons de qualité supérieure, afin de réduire le débit binaire pour coder les qualités supérieures.

La scalabilité SNR s'articule essentiellement sur l'adoption de paramètres de quantification (*Quantization Parameter* QP) distincts pour chaque couche. La norme H.264/SVC prend en charge deux modes distincts de scalabilité SNR :

- *Coarse Grain Scalability* (CGS) : dans ce mode, chaque couche a une procédure de prédiction indépendante (toutes les références ont le même niveau de qualité). En fait, le mode CGS peut être considéré comme un cas particulier de scalabilité spatiale quand des couches consécutives ont la même résolution.

- *Medium Grain Scalability* (MGS) : l'approche MGS augmente l'efficacité en utilisant un module de prédiction plus souple, où les deux types de couches (de base et d'amélioration) peuvent être référencés. Cependant, cette stratégie peut induire un décalage de synchronisation entre

le codeur et le décodeur quand il y a un retard de transmission de la couche d'amélioration. Pour résoudre ce problème, la spécification MGS propose l'utilisation d'images clés périodiques, qui resynchronise immédiatement le module de prédiction.

La Figure 2.5 montre la différence entre ces deux modes de scalabilité SNR.

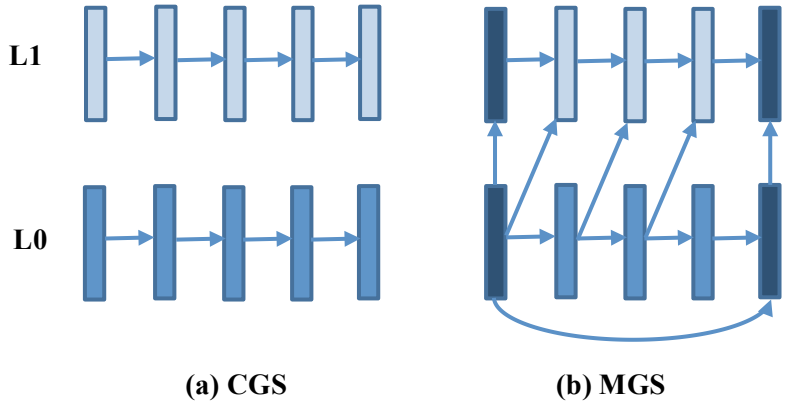


Figure 2.5. Différence entre CGS et MGS

SVC maintient l'organisation du *bitstream* introduit en H.264/AVC. Plus précisément, tous les composants du *bitstream* sont encapsulés dans des unités de la couche d'abstraction de réseau (NALU). Un sous-ensemble des NALU correspond à la couche de codage vidéo (VCL), et contient les données d'images codées associées au contenu de la source.

A ce jour, le déploiement de SVC pour le streaming multimédia et en particulier IPTV est un domaine de recherche actif. Pour avoir plus de détails là-dessus, veuillez vous référer aux références suivantes : [10][11][12][13][14][15].

2.4 Les paramètres qui influencent la qualité d'expérience (QoE) de l'IPTV

Il y a beaucoup de facteurs qui ont un impact sur la qualité vidéo perçue. Ces facteurs dépendent de l'application, de la technologie des réseaux, du terminal de l'utilisateur etc. Il est possible de classer les facteurs qui affectent la qualité en quatre catégories, en fonction de la source du facteur :

- **Paramètres environnement**. Les paramètres environnement sont par exemple : l'éclairage de la pièce, la fidélité de l'écran (moniteur ou TV), les capacités de calcul du lecteur multimédia (l'ordinateur), etc. Les paramètres environnement sont généralement incontrôlables et difficiles à mesurer dans une session de test.

- **Paramètres de la source**. Le signal source de la vidéo a un impact évident et fort sur la qualité globale perçue. Par exemple, le niveau de luminance et la quantité de mouvement des scènes ont un impact important sur la qualité, en particulier quand il y a d'autres facteurs, comme l'encodage à débit binaire très faible et/ou des pertes de paquets dans le réseau. Les paramètres de la source qui dépendent des caractéristiques de la séquence, comme la nature de la scène (par exemple, la quantité de mouvements, la couleur, le contraste, la taille de l'image, etc.) ont également un impact sur la perception humaine de la qualité de la vidéo. L'encodage ou les paramètres de compression sont les facteurs sources les plus importants.

Nous pouvons citer comme exemples de ces paramètres : le type de codec utilisé (MPEG-2, MPEG-4 partie 2 ou 10, etc.), le nombre de bits par échantillon, le débit binaire, la fréquence d'images, le nombre de couches dans le cas de codage en couches, etc. L'expéditeur de la vidéo peut mettre en œuvre des techniques d'amélioration de la qualité (le récepteur peut également interpréter les données ajoutées ou modifiées). Les plus courants sont : *Forward Error Correction* (FEC), l'entrelacement et le codage en couches. Généralement, ces techniques d'amélioration sont conçues pour atténuer l'effet des pertes de paquets dans le réseau sur la qualité.

- **Paramètres réseau**. La qualité de service (QoS) est un élément principal dans la conception du réseau et de la gestion en général. En règle générale, les paramètres QoS comprennent la perte de paquets, le délai, la gigue et les facteurs de bande passante. L'effet de ces paramètres sur la qualité perçue dépend essentiellement du type de l'application multimédia. Par exemple, si un service temps réel est nécessaire, le taux de perte de paquets est le paramètre réseau le plus important. La retransmission et la mise en mémoire tampon jouent un rôle important aussi. S'il y a une interactivité (par exemple lors d'un appel vidéo), le délai et la gigue ont également un rôle important, en ajoutant l'écho et la perte de synchronisation audio/vidéo.
- **Paramètres du récepteur**. Outre les techniques d'amélioration de la qualité, mis en place par l'expéditeur, il existe un ensemble de procédures, améliorant la qualité, qui peuvent être mises en œuvre au niveau du récepteur. Voici quelques exemples : la mise en mémoire tampon, la dissimulation de perte (d'insertion, d'interpolation et de régénération des données perdues) et des améliorations de contrôle de congestion dans les flux UDP.

Dans la section suivante, nous présentons des méthodes qui permettent d'évaluer la qualité des séquences vidéo.

2.5 Evaluation de la qualité vidéo

Une bonne méthode d'évaluation de la qualité peut aider à contrôler la qualité des services vidéo et le renforcement de l'expérience de l'utilisateur. En raison de son rôle fondamental, un grand nombre de mesures objectives de la qualité de la vidéo ont été proposées au fil du temps.

Les méthodes à base de pixels, telles que l'erreur quadratique moyenne (*Mean Square Error* MSE), *Signal-to-Noise Ratio* (SNR) et *Peak Signal-to-Noise Ratio* (PSNR), sont simples à calculer et faciles à intégrer dans le processus d'optimisation. Toutefois, il est bien reconnu que ces mesures à base de pixels ne correspondent pas bien à la perception du système visuel humain. Les distorsions perçues par l'être humain ne sont pas toujours prises en compte par MSE/SNR/PSNR, parce que ces mesures fonctionnent sur une base « pixel par pixel » sans tenir compte du contenu du signal, la condition de visualisation et les caractéristiques du système visuel humain (HVS). Ces problèmes rendent nécessaire la conception d'une meilleure métrique objective de qualité.

A la différence des évaluations objectives de la qualité de la vidéo, qui fonctionnent d'une façon automatique, sans intervention humaine, l'évaluation subjective est basée sur la qualité de jugement des observateurs humains. L'évaluation subjective est considérée comme la méthode la plus précise pour mesurer la qualité visuelle. Bien que l'évaluation subjective de la qualité prenne du temps et ne soit pas

réalisable en temps-réel, son rôle dans la conception de métrique objective de la qualité est encore irremplaçable : la qualité de perception visuelle dérivée de l'évaluation subjective peut servir de référence pour l'évaluation de la performance des évaluations objectives, et peut même être un repère pour la conception de nouvelles métriques objectives.

Dans cette section, nous présentons les problèmes liés à l'évaluation de la qualité des vidéos. Nous discuterons des méthodes d'évaluation actuellement disponibles dans la littérature et pourquoi elles ne répondent pas nécessairement aux besoins actuels en matière d'évaluation de la qualité perçue.

2.5.1 Evaluation subjective

Les évaluations subjectives représentent la façon la plus précise pour mesurer la qualité d'une vidéo. Dans les expériences subjectives, un certain nombre de sujets (observateurs ou participants) sont invités à assister à un ensemble de tests et de donner un jugement sur la qualité des vidéos ou l'incommodité causée par les distorsions. La moyenne des valeurs obtenues pour chaque séquence de test est connue sous le nom *Mean Opinion Score* (MOS).

En général, les évaluations subjectives sont coûteuses et nécessitent beaucoup de temps. En conséquence, le nombre d'expériences qui peuvent être réalisées est limité et, par suite, une méthodologie appropriée doit être utilisée pour tirer le meilleur parti des ressources. L'Union Internationale des Télécommunications (*International Telecommunication Union* ITU) a formulé des recommandations pour les procédures des tests subjectifs. Les deux documents les plus importants sont la Recommandation ITU-R. BT.500-11 [16], destinée aux applications de télévision, et la Recommandation ITU-T. P.910 [17], destinée aux applications multimédia. Ces documents donnent des informations sur les conditions d'affichage des vidéos, les critères de sélection des observateurs et du matériel d'essai, les procédures d'évaluation, et les méthodes d'analyse des données. Avant de choisir la méthode à utiliser, l'expérimentateur doit tenir compte de l'application et des objectifs de l'évaluation.

Selon l'ITU, il existe deux catégories d'évaluations subjectives :

- Évaluations de la qualité : les notes rendues par les participants sont sur une échelle de qualité, c'est-à-dire, la qualité de la vidéo affichée est-elle bonne ou mauvaise. Ces évaluations sont utilisées pour évaluer la performance des systèmes utilisés dans des conditions optimales.
- Les tests de dépréciation : les jugements rendus par les sujets sont sur une échelle de valeur, c'est-à-dire, les distorsions de la vidéo affichée sont-elles visibles ou imperceptibles. Ces évaluations sont utilisées pour évaluer la capacité des systèmes à conserver la qualité des vidéos dans des conditions non optimales. Ces méthodes sont souvent utilisées pour mesurer la dégradation de la qualité causée par un codage ou un schéma de transmission.

Les échelles d'évaluation, pour l'évaluation de la qualité ou pour l'évaluation des dépréciations, peuvent être continues ou discrètes. Les jugements peuvent également être catégoriques ou non catégoriques, adjectivaux ou numériques. En fonction de la forme de présentation de la séquence vidéo (stimulus), les méthodes d'évaluation peuvent être classées en tant que stimulus simple ou double. Dans l'approche du stimulus simple, seule la séquence de test est présentée, tandis que dans la méthode double

stimulus, une paire de séquences (séquence de test et la séquence de référence correspondante) sont présentées ensemble.

Parmi les procédures de tests subjectifs, proposées dans ITU-R Rec. BT.500-11, nous pouvons citer :

- *Double Stimulus Impairment Scale* (DSIS) – Pour cette méthode, la séquence de référence est toujours affichée avant la séquence de test et la paire ne se répète pas. Les observateurs sont invités à juger le niveau des dépréciations pour chaque séquence de test, en utilisant une échelle à cinq niveaux. Cette méthode est appropriée pour évaluer les artefacts visibles.

- *Double Stimulus Continuous Quality Scale* (DSCQS) – Dans cette méthode, des paires de séquences multiples (contenant la référence et une séquence dégradée au hasard) sont présentées aux observateurs. DSCQS est utile quand il n'est pas possible de fournir les conditions d'essai qui montrent la gamme complète de la qualité.

- *Single Stimulus Continuous Quality Evaluation* (SSCQE) – Dans cette méthode, les observateurs sont invités à regarder une vidéo d'environ 20-30 minutes. La séquence de référence n'est pas présentée. L'observateur utilise continuellement un curseur pour évaluer la qualité, puisqu'elle change au cours de la présentation. SSCQE est conçue pour mesurer la qualité dans un contexte variable, par exemple, quand une compression adaptative est utilisée.

Les procédures d'évaluation les plus populaires de l'ITU-T Rec. P.910 sont les suivantes :

- Absolute Category Rating (ACR) - Aussi connu comme méthode de stimulation simple (*Single Stimulus Method* SSM), ce procédé est caractérisé par le fait que les séquences de test sont présentées une à une, sans la séquence de référence. Cela en fait une méthode très efficace, par rapport au DSIS ou DSCQS, qui ont des durées de l'ordre de 2 à 4 fois plus longues. Après chaque présentation, les observateurs sont invités à juger la qualité globale de la séquence de test en utilisant une échelle à cinq niveaux. Une échelle de neuf niveaux peut être utilisée si l'expérimentateur nécessite une distinction supplémentaire entre les jugements.

- *Degradation Category Rating* (DCR) - Cette méthode est identique à la DSIS décrite précédemment.

- *Pair Comparison* (PC) - Dans cette méthode, toutes les combinaisons de paires possibles (vidéo originale et vidéo altérée) de toutes les séquences de test sont présentées aux observateurs. Les observateurs doivent choisir la séquence de la paire qui a la meilleure qualité. Cette méthode permet une distinction très fine entre les états des vidéos, mais exige également une plus longue période de temps par rapport à d'autres méthodes.

Les différences entre ces procédures sont minimes et dépendent principalement de l'application envisagée. Elles concernent, par exemple, le fait que dans les tests subjectifs, on montre au préalable aux observateurs les séquences de référence, l'échelle d'évaluation de la qualité (et le fait qu'elle est discrète ou continue, voir Figure 2.6), la longueur de la séquence (généralement une dizaine de secondes), le nombre de vidéos par essai (une fois, deux fois de suite ou deux fois simultanément), la possibilité de modifier les valeurs données précédemment ou non, etc.

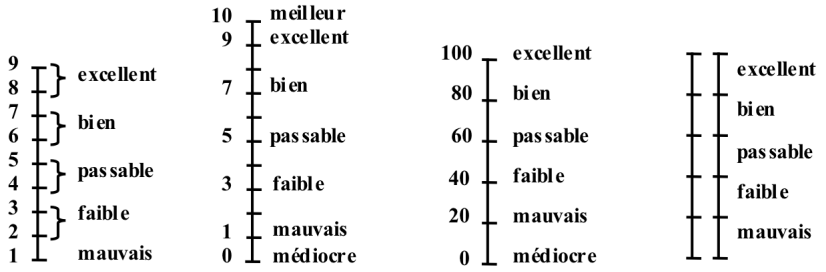


Figure 2.6. Exemples d'échelles pour les méthodes d'évaluation subjective

D'autres méthodes d'évaluation subjective sont disponibles, comme : *Simultaneous Double Stimulus for Continuous Evaluation* (SDSCE), *Stimulus Comparison Adjectival Categorical Judgement* (SCACJ) et *Subjective Assessment Methodology for VIdeo Quality* (SAMVIQ). Pour plus de détails sur ces tests subjectifs, veuillez vous référer à [16][17].

Toutes les méthodes présentées sont des méthodes d'évaluation subjective de la qualité qui mesurent la qualité perçue du point de vue de l'utilisateur. Cependant, les tests subjectifs sont très longs, fastidieux et coûteux en main-d'œuvre, ce qui les rend difficiles à automatiser et à répéter. En outre, compte tenu de leur nature, ces méthodes ne sont évidemment pas appropriées pour un fonctionnement en temps réel. Pour ces raisons, beaucoup d'efforts ont été concentrés sur le développement des méthodes d'évaluation objective moins coûteuses, plus rapides et plus faciles à utiliser.

2.5.2 Evaluation objective

Les méthodes objectives sont des algorithmes et des formules (généralement des algorithmes de traitement de signal) qui mesurent, d'une certaine manière, la qualité d'une vidéo. Les mesures objectives de la qualité vidéo peuvent aller du très simple au très complexe. En particulier, celles qui sont fondées sur les systèmes de vision de l'homme (*Human Vision Systems* HVS) ont tendance à être très complexes, de sorte que même si la version originale et les versions dégradées sont disponibles, l'évaluation ne peut pas se faire en temps réel.

Les méthodes d'évaluation objective de la qualité vidéo peuvent être classées en trois catégories sur la base de la quantité d'informations disponibles pour la comparaison avec la vidéo d'origine :

- *Full-Reference* (FR) : La vidéo originale et la vidéo déformée sont disponibles.
- *Reduced-Reference* (RR) : En plus de la vidéo altérée, une description de la vidéo originale et certaines de ses caractéristiques sont disponibles.
- *No-Reference* (ou *Free-Reference*) (NR) : Seule la vidéo altérée est disponible.

La Figure 2.7 représente les schémas correspondant aux méthodes objectives d'évaluation de la qualité : *Full-Reference*, *Reduced-Reference* et *No-Reference*. Nous remarquons qu'avec l'approche FR, la vidéo de référence est disponible en entier au niveau du point de mesure. Dans l'approche RR, une partie de la vidéo de référence est disponible et transmise à travers un canal auxiliaire. Dans ce cas, les informations disponibles au point de mesure sont généralement constituées d'un ensemble de caractéristiques extraites

de la vidéo de référence. Pour l'approche NR, aucune information concernant la vidéo de référence n'est disponible au point de mesure.

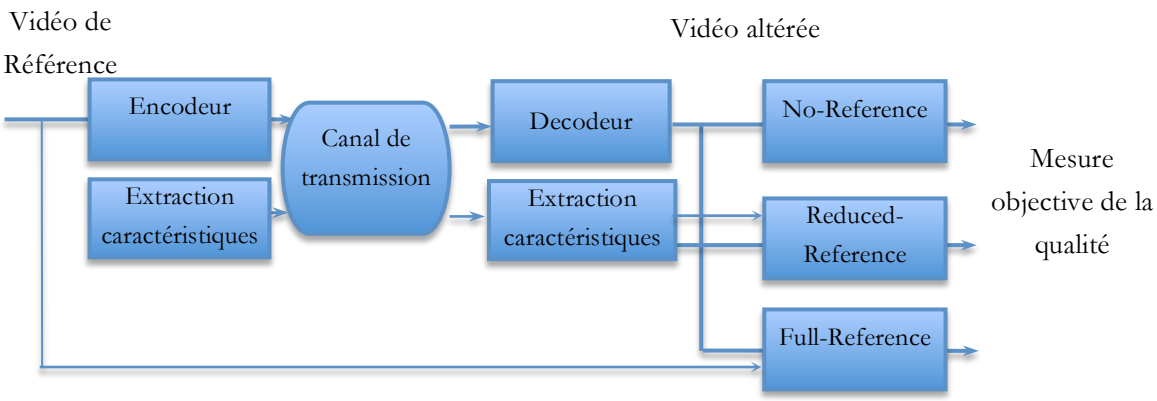


Figure 2.7. Diagramme des méthodes d'évaluation objectives : Full-Reference, Reduced-Reference et No-Reference

Ces trois catégories de mesure sont ciblées pour des applications différentes. Les métriques FR sont plus appropriées pour mesurer la qualité en hors ligne, où une mesure précise et détaillée de la qualité de la vidéo est plus importante que d'avoir des résultats immédiats. Les mesures NR et RR sont destinées aux applications en temps réel, où les limites de complexité de calcul et le manque d'accès à la séquence de référence sont les principaux problèmes.

Dans la section suivante, une brève description de quelques méthodes d'évaluation objective (FR, RR et NR) est présentée.

2.5.2.1 Métriques *Full-Reference*

La plupart des mesures de qualité proposées dans la littérature sont des mesures FR, c'est-à-dire les vidéos originales et déformées doivent être disponibles pour mesurer la qualité. Ces mesures estiment la qualité d'une vidéo en comparant la vidéo de référence et la vidéo altérée.

2.5.2.1.1 Mean Square Error (MSE) et Peak Signal–to–Noise Ratio (PSNR)

MSE et PSNR sont les métriques FR les plus utilisées. Si l'on considère un clip vidéo comprenant K images de $M \times N$ pixels chacune, nous pouvons définir l'erreur quadratique moyenne (MSE) comme suit :

$$MSE = \frac{1}{K \cdot M \cdot N} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} [o_k(m,n) - d_k(m,n)]^2 \qquad (2.1)$$

avec $o_k(m,n)$ et $d_k(m,n)$ les pixels de position (m,n) dans la $k^{ième}$ image de la séquence originale et dégradée.

Le PSNR est défini comme suit :

$$PSNR = 10log_{10}\left(\frac{max_k}{MSE}\right) \qquad (2.2)$$

avec max_k étant la valeur maximale dans l'image k.

Le MSE mesure la différence entre les images alors que le PSNR mesure la fidélité des images, c'est-à-dire à quel point deux images se ressemblent. Le MSE et le PSNR sont souvent utilisés en raison de leurs significations physiques et de leur simplicité. Cependant les résultats de ces mesures sont assez pauvres, car ils ne fournissent généralement pas une bonne corrélation avec les scores subjectifs.

2.5.2.1.2 Structural SIMilarity (SSIM)

La métrique SSIM mesure la similarité structurelle en se basant sur le modèle HVS. Cette méthode utilise la mesure de la distorsion structurelle à la place de l'erreur. SSIM est basée sur le fait que le HVS est plus sensible aux changements structurels des vidéos qu'à ceux de luminance et de contraste. SSIM estime la qualité de la vidéo en extrayant, à partir des images, des informations telles que des informations de structure et de contraste, et en comparant les valeurs de ces informations au lieu de comparer directement les pixels. Des études sur la performance du SSIM ont montré que cette métrique simple offre de bons résultats [18].

2.5.2.1.3 NTIA Video Quality Metric (VQM)

VQM [19] a été développée par l'Institut des sciences de télécommunication de *Boulder Colorado*. VQM est une méthode normalisée pour la mesure objective de la qualité vidéo. VQM fait une comparaison entre la séquence vidéo originale et les séquences vidéo présentant une distorsion, basées uniquement sur un ensemble de caractéristiques extraites indépendamment de chaque vidéo.

L'algorithme utilisé par VQM mesure les effets perceptifs de plusieurs distorsions, telles que le flou, le mouvement saccadé/non-naturel, le bruit, la distorsion des blocs et la distorsion des couleurs. Ces mesures sont combinées en une seule mesure qui donne une prédiction de la qualité globale.

VQM prend la vidéo originale et la vidéo traitée en entrée et estime la qualité comme suit :

1. Calibration : cette étape calibre la vidéo altérée pour l'extraction de certaines caractéristiques. VQM estime et corrige le décalage spatial et temporel ainsi que le décalage du contraste et de la luminosité de la séquence vidéo traitée par rapport à la séquence vidéo d'origine.
2. Extraction des caractéristiques de qualité : cette étape extrait, à l'aide d'une fonction mathématique, un ensemble de caractéristiques de la qualité qui caractérise les changements de perception dans les propriétés spatiales, temporelles et de chrominance à partir des sous-régions spatio-temporelles de la vidéo.
3. Calcul des paramètres de qualité : cette étape calcule un ensemble de paramètres de qualité qui décrivent des changements dans la qualité de la vidéo en comparant les caractéristiques extraites de la vidéo traitée avec celles extraites de la vidéo d'origine.
4. Estimation de la qualité : il s'agit de la dernière étape. VQM calcule la qualité globale de la vidéo en utilisant une combinaison linéaire des paramètres calculés à partir des étapes précédentes.

Le modèle d'estimation de la qualité est normalisé par l'ITU-R BT.1683 [20], et des études sur ses performances ont montré qu'il a une forte corrélation avec les scores des évaluations subjectives [21].

Le fait que ces méthodes nécessitent la vidéo de référence limite leurs utilisations dans des applications de services vidéo, où les vidéos originales ne sont pas accessibles. En outre, un alignement spatial et temporel très précis, entre la vidéo originale et la vidéo traitée, est nécessaire pour garantir la précision de certains indicateurs.

Pour ces raisons, des métriques RR ont été développées pour permettre l'évaluation de la qualité perçue, en se basant sur des caractéristiques structurelles extraites de la vidéo d'origine et de la vidéo altérée.

2.5.2.2 Métriques *Reduced-Reference*

Les mesures de qualité vidéo RR nécessitent uniquement des informations partielles sur la vidéo de référence. Pour pouvoir évaluer la qualité de la vidéo, certaines caractéristiques physiques sont extraites de la référence et transmises au récepteur. L'une des caractéristiques intéressantes de mesures RR est la possibilité de choisir la quantité d'informations à extraire et à envoyer. Dans la pratique, la quantité d'informations à envoyer dépend des caractéristiques du canal utilisé pour transmettre les données ou du stockage disponible pour les mettre en cache.

Les différentes métriques de cette classe peuvent être moins précises que les mesures de référence FR, mais elles sont moins complexes, et leurs implémentations pour une utilisation en temps réel sont plus abordables. Néanmoins, la synchronisation entre les données originales et les données traitées est toujours nécessaire.

On peut citer comme exemple de métrique RR : l'algorithme de *Webster* [22], *Local Harmonic Strength* (LHS) [23] et l'algorithme proposé par *Lin Ma* [24].

2.5.2.2.1 Méthode proposée par Webster et al.

Cette métrique RR est une métrique d'extraction de caractéristiques qui estime le montant de la distorsion dans une vidéo par l'extraction des activités spatiales (*Spatial Information* SI) et temporelles (*Temporal Information* TI) en utilisant des filtres spécialement conçus.

L'information temporelle (TI) correspond à l'écart type de la différence de trames, c'est-à-dire, la quantité de mouvements dans la vidéo. Trois paramètres de comparaison sont dérivés des caractéristiques SI et TI de la vidéo de référence et des vidéos déformées. Les paramètres de la vidéo de référence sont transmis sur le canal. La taille des données RR dépend de la taille de la fenêtre sur laquelle les caractéristiques SI et TI sont calculées.

On rappelle que, selon [17], les quantités d'informations spatiale et temporelle sont utilisées pour caractériser une séquence vidéo. La mesure de l'information spatiale (SI) est basée sur le filtre *Sobel*, qui est appliqué à chaque trame de luminance F_n à l'instant n. Ensuite, l'écart-type sur les pixels est calculé. La valeur maximale dans la séquence représente la totalité des informations spatiales :

$$SI = max_{temps}\{std_{space}[Sobel(F_n)]\} \quad \textbf{(2.3)}$$

La mesure de l'information temporelle (TI) est basée sur la différence de mouvements. Pour tout instant n, la différence des valeurs de luminance d'un pixel est calculée :

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \quad \textbf{(2.4)}$$

TI est calculé, comme dans le cas de SI, par la formule suivante :

$$TI = max_{temps}\{std_{space}[Sobel(M_n(i,j))]\} \quad \textbf{(2.5)}$$

2.5.2.2.2 Méthode proposée par Gunawan and Ghanbari

Gunawan et Ghanbari proposent une métrique RR qui est basée sur le gain/perte harmonique d'informations par le biais d'une analyse discriminante de force harmonique locale (*Local Harmonic Strength* LHS). La force harmonique peut être interprétée comme une mesure de l'activité spatiale. La LHS est calculée à partir du gradient de l'image et sa valeur représente un degré relatif de l'effet de bloc sur l'image. En outre, la comparaison entre les valeurs de la LHS, à partir d'une vidéo originale et d'une vidéo traitée, montre que la LHS peut également être utilisée pour indiquer d'autres types de dégradations, telles que l'effet de flou qui correspond à une perte d'énergie.

Dans un algorithme plus récent, les auteurs ont amélioré les performances de l'algorithme en comprimant les informations RR [25], ce qui induit une réduction de la quantité de données RR qui doivent être transmises ou stockées.

2.5.2.2.3 Méthode proposée par Lin Ma

La méthode RR d'évaluation de la qualité vidéo proposée par *Lin Ma*, exploite les caractéristiques temporelles statiques de l'histogramme inter-trame et les pertes d'informations spatiales. Du point de vue spatial, l'auteur a proposé un descripteur de variation d'énergie (*Energy Variation Descriptor* EVD) pour mesurer la variation d'énergie de chaque image codée, qui résulte de l'opération de quantification. En plus de représenter la variation d'énergie, EVD peut encore simuler la propriété de masquage de texture du système visuel humain (HVS). Du point de vue temporel, la densité gaussienne généralisée (*Generalized Gaussian Density* GGD) est utilisée pour capturer les statistiques naturelles de la distribution d'histogramme inter-trame. La distance city-bloc (*City-Block Distance* CDB) est utilisée pour calculer la distance entre l'histogramme de la séquence vidéo originale et celle codée. La mesure de la qualité de vidéo est obtenue en combinant l'EVD spatiale avec la CDB temporelle.

2.5.2.3 Métriques *No-Reference*

La nécessité et la disponibilité des vidéos de référence ou même d'une petite partie de celles-ci deviennent un obstacle sérieux dans de nombreuses applications de transmission en temps réel. Il devient essentiel de trouver des moyens pour estimer la qualité d'une vidéo à l'aide d'une métrique NR. Bien que les observateurs humains puissent généralement évaluer la qualité d'une vidéo sans utiliser la vidéo de référence, la conception d'une métrique sans référence est une tâche très difficile. La plupart des métriques NR proposées dans la littérature sont des mesures d'extraction de caractéristiques qui permettent d'estimer la qualité de la vidéo. Pour des raisons de complexité, plusieurs mesures NR s'appuient sur une ou deux caractéristiques pour estimer la qualité. Dans la plupart des cas, les caractéristiques utilisées dans les algorithmes d'estimation de la qualité vidéo sont des artefacts, tels que le blocage (*Blockiness*), le flou (*Blurriness*), etc.

2.5.2.3.1 Méthode proposée par bSoft

La méthode proposée par l'entreprise bSoft [26] est une méthode sans référence (NR), qui essaye d'estimer le MOS en extrayant des paramètres vidéo et réseau, tels que le débit binaire '*b*', taux d'image (*frame rate*) '*f*', taux de perte de paquets '*l*', blocage de vidéo '*j*', paramètre de quantification '*q*', etc.

$$MOS = B \times b^{bb} + F \times f^{ff} + L \times l^{ll} + J \times j^{jj} + \cdots + Q \times q^{qq} \quad \textbf{(2.6)}$$

Ces paramètres sont combinés par une formule non linéaire, comme indiqué dans l'expression (2.6), en utilisant des facteurs de pondération et des exposants spécifiques, qui peuvent être ajustés en fonction du service et en fonction du terminal utilisé par l'application.

2.5.2.3.2 Méthode proposée par Demokritos

Cette méthode, proposée par l'équipe Demokritos [27], considère la corrélation entre les taux de perte de paquets réseau, la taille des paquets et le taux d'image décodables théoriquement prévu, noté *Q*. *Q* est une mesure de niveau applicative, définie comme la fraction des trames décodables, qui est le nombre de trames théoriquement décodables attendues et du nombre total de trames émises par la source vidéo :

$$Q = \frac{N_{dec}}{(N_{total-I} + N_{total-P} + N_{total-B})} \quad \textbf{(2.7)}$$

N_{Dec} est la somme de plusieurs trames *I, P, B* théoriquement décodées avec succès, c'est-à-dire $N_{dec\text{-}I}$, $N_{dec\text{-}P}$, et $N_{dec\text{-}B}$.

Compte tenu de la structure GOP et en prenant en considération les interdépendances du décodage entre les trois types de trames, l'impact du taux de perte de paquets peut être mathématiquement formulé. La correspondance entre les trames perdues et le MOS dans le cas de vidéo, avec un contenu spatio-temporel varié, d'une durée de 10 secondes, et avec une résolution CIF 25fps, est analytiquement décrite par l'expression suivante :

$$PQoS\ Level = 85{,}5 - \frac{53{,}03}{1 + \left(\frac{562}{(1-Q)\cdot 10^4}\right)^{1{,}01}} \quad \textbf{(2.8)}$$

Par exemple, pour le cas de zéro perte de trames, la formule donne une note de 85,8 sur 100, ce qui correspond à «Excellent» en utilisant l'échelle de notation « *Double Stimulus using a Continuous Quality Scale* ».

2.5.2.3.3 Méthode PSQA

Pseudo-Subjective Quality Assessment est une mesure non-paramétrique sans référence de la QoE. PSQA est basée sur un modèle mathématique, appelé les réseaux de neurones. Nous détaillerons la méthode PSQA dans le chapitre 4.

Conclusion

Dans ce chapitre, nous avons présenté le principe de codage et les problèmes liés à l'évaluation de la qualité des vidéos. Nous avons examiné à la fois les méthodes d'évaluations objective et subjective de la qualité visuelle.

La première partie de ce chapitre décrit brièvement les différents aspects de codage vidéo. Nous avons présenté les différents codecs MPEG les plus utilisée pour la vidéo sur IP. Nous avons ensuite présenté les procédures d'évaluation subjective et les méthodes objectives les plus connues dans la littérature. Ces méthodes présentées ont beaucoup de limitations et de contraintes d'utilisation. Nous proposerons dans le chapitre 6 nos propres méthodes pour l'évaluation de la qualité vidéo.

Chapitre 3

3 La Voix sur IP

Les applications VoIP permettent d'émettre et de recevoir des appels vocaux. Une application VoIP (par exemple, Skype, Google Talk, …) fournit normalement de nombreux codecs vocaux qui peuvent être sélectionnés ou mis à jour manuellement ou automatiquement. Les codecs vocaux typiques utilisés dans la VoIP comprennent ceux proposés par l'ITU-T telles que G.711, G.729 et G.723.1 ; par ETSI tels que AMR; les codecs open-source tels que les codecs iLBC et Speex ; et les propriétaires tels que le codec Silk de Skype. Ces codecs ont un débit variable dans la gamme de 6 à 40kbit/s et une fréquence d'échantillonnage variable sur une bande étroite à une bande super large. Certains codecs ne peuvent fonctionner qu'à un débit binaire fixe, tandis que de nombreux codecs avancés peuvent avoir des débits binaires variables qui peuvent être utilisées pour l'adaptation afin d'améliorer la qualité de la voix ou la qualité d'expérience.

Les codecs vocaux sont basés sur différentes techniques de compression de la parole visant à éliminer la redondance dans le signal de parole pour obtenir une bonne compression et de réduire les coûts de stockage et de transmission. Les codecs peuvent atteindre des taux de compression de 53,3% de la séquence vocale, tout en conservant l'intelligibilité, mais avec une qualité vocale qui est un peu moins bonne. La plupart des codecs vocaux fonctionnent dans la gamme de 4,8 kbit/s à 16 kbit/s et ont une qualité vocale et un taux de compression raisonnable. Ces codecs sont principalement utilisés dans des applications mobiles/sans fil, là où les ressources en bande passante sont limitées. En général, plus le débit binaire de la parole est grand, plus la qualité de la parole est bonne et plus l'application est gourmande en bande passante et en stockage. Dans la pratique, c'est toujours un compromis entre l'utilisation de la bande passante et la qualité vocale.

Dans ce chapitre, nous commençons par présenter brièvement les bases de la compression de la parole. Nous présentons ensuite les codecs les plus utilisés par les applications VoIP. Dans la troisième partie de ce chapitre, nous expliquons les différents aspects qui peuvent altérés la qualité d'une communication VoIP. Enfin nous présentons les différentes méthodes d'évaluation (subjectives et objectives) de la qualité des communications VoIP.

3.1 Acquisition et reconstruction de la voix

Avant d'envoyer des informations voix sur le réseau, nous devons d'abord numériser le signal analogique de la voix. La numérisation se réalise en deux étapes, l'échantillonnage et la quantification. Elle va permettre de transformer un signal continu en une suite de valeurs discrètes qui seront traduites dans le langage des ordinateurs, en 0 et 1.

Après la transmission, le destinataire du signal numérique doit le convertir en analogique, pour générer un son à la sortie des haut-parleurs. Ces étapes sont appelées aussi une conversion analogique-à-numérique (analogue-to-digital A/D) et numérique-à-analogique (digital-to-analogue D/A).

Il est utile de savoir que la numérisation d'un signal audio est fréquemment désignée par *Modulation par Impulsions Codées* (MIC) ou *Pulse Code Modulation* (PCM) en anglais.

3.1.1 Echantillonnage et quantification

Un signal continu (signal voix par exemple), sur un certain intervalle de temps, a une infinité de valeurs, avec une précision infinie. Afin de pouvoir enregistrer une approximation numérique du signal, il est tout d'abord échantillonné, puis quantifié, comme le montre la Figure 3.1.

L'échantillonnage est la première phase de la numérisation qui consiste à passer d'un signal continu en une suite de valeurs mesurées à intervalles réguliers. Le signal analogique est ainsi découpé en "tranches" ou échantillons (*samples*). Le nombre d'échantillons audio par seconde représente la fréquence d'échantillonnage ou *Sampling Rate*. Celle-ci est exprimée en *Hertz* (Hz).

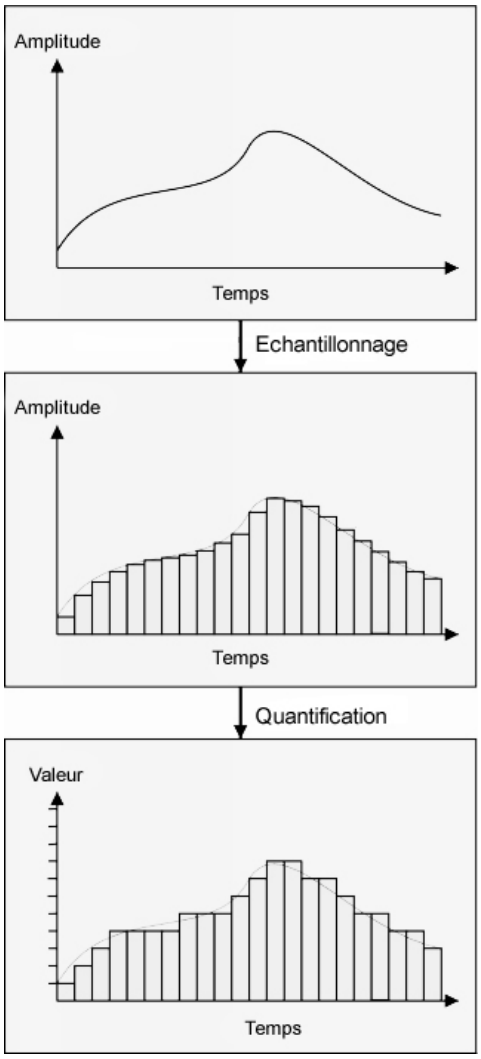


Figure 3.1. Echantillonnage et Quantification d'un signal analogique

La quantification est la seconde étape de la numérisation. Après avoir découpé le signal continu en échantillons, il va falloir les mesurer et leur donner une valeur numérique en fonction de leur amplitude. Pour cela, nous définissons un intervalle de *N* valeurs destiné à couvrir l'ensemble des valeurs possibles. Ce nombre *N* est codé en binaire sur 8, 16, 20 ou 24 bits suivant la résolution du convertisseur A/N. L'amplitude de chaque échantillon est alors représentée par un nombre entier.

3.1.2 Reconstruction du signal

La reconstruction d'un signal fait l'inverse de l'étape de numérisation. Une quantification inverse est appliquée et à partir de ces échantillons, un signal continu est recréé. La ressemblance entre le signal reconstruit et le signal d'origine dépend de la fréquence d'échantillonnage, du procédé de quantification et de l'algorithme de reconstruction utilisé. Une bonne introduction, sur la théorie de la reconstruction d'un signal, peut être trouvée dans [28].

3.2 Les codecs pour la VoIP

3.2.1 ITU G.711

G.711 [29] est un codec qui a été mis en place par l'ITU en 1972 pour la téléphonie numérique. Le codec a deux variantes : A-Law est utilisé en Europe et lors des communications internationales, μ-Law est utilisé dans les États-Unis d'Amérique et le Japon.

G.711 utilise une compression logarithmique. Le codec transforme les échantillons de 16-bit en 8 bits, il atteint un taux de compression de 50%. Le débit résultant, pour une seule direction, est de 64 kbit/s, donc un appel consomme 128 kbit/s. Ce codec peut être librement utilisé (*open-source*) dans des applications Voix sur IP. Les meilleures performances de ce codec sont obtenues dans les réseaux locaux où nous avons beaucoup de bande passante disponible. De plus, ce codec est caractérisé par une très bonne qualité audio perçue (un MOS de 4,2 sur 5) et par une simple implémentation, et donc il n'a pas besoin d'un processeur puissant.

3.2.2 ITU G.729

Le standard G.729 [30] décrit un algorithme pour le codage de signaux vocaux à 8 kbit/s au moyen de la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP) (*Conjugate-Structure Algebraic-Code-Excited Linear-Prediction*).

Les annexes A, B et D à J du standard G.729 étendent les fonctionnalités du codec tel que le codage à un taux de 6,4 kbit/s et 11,8 kbit/s, une multi-cadence de fonctionnement, DTX et fournit une version à complexité réduite de l'algorithme. Le codec G.729 est généralement utilisé dans les applications VoIP.

3.2.3 ITU G.723.1

Le codec G.723.1 [31] est aussi généralement utilisé dans les applications VoIP. Le codeur est fondé sur les principes du codage prédictif linéaire (*Linear Predictive Coding* LPC) par analyse et synthèse, en vue de minimiser un signal d'erreur pondéré par une courbe de perception.

Le G.723.1 possède deux débits binaires associés : 5,3 kbit/s et 6,3 kbit/s. Pour le débit supérieur, on fait appel à l'excitation par quantification d'impulsions multiples selon le critère du maximum de vraisemblance (MP-MLQ, *Multi-Pulse Maximum Likelihood Quantization*). Pour le débit inférieur, on fait appel à l'excitation par séquences codées à structure algébrique (ACELP, *Algebraic-Code-Excitation*).

3.2.4 GSM-FR

Le codec *GSM 06.10 Full-Rate* [32] décrit le transcodage dans la télécommunication cellulaire. Le schéma de codage est basé sur *Regular Pulse Excitation – Long Term Prediction* (RPE-LTP) paradigme de codage de la parole.

3.2.5 GSM-HR

GSM 06.20 Half Rate (HR) [33] nécessite moins de la moitié de la bande passante du GSM-FR au prix d'une moins bonne qualité audio. Le codec utilise l'algorithme *Vector-Sum Excited Linear Prediction* (VSELP).

3.2.6 AMR

Le codec audio *Adaptive Multi-Rate* (AMR) [34] est largement utilisé dans les réseaux cellulaires GSM et UMTS. Le codec encode le signal à huit différents débits, de l'ordre de 4,75 kbit/s à 12,2 kbit/s. Le débit le plus élevé de 12,2 kb/s est compatible avec le standard *GSM Enhanced Full Rate* (GSM-EFR). Le système de codage est basé sur l'algorithme *Algebraic Code Excited Linear Prediction* (ACELP).

3.2.7 iLBC

Internet Low Bitrate Codec (iLBC) [35] est un codec conçu pour la communication voix sur IP. L'algorithme utilise *Block-Independent Linear Predictive Coding* (BI-LPC) et prend en charge des débits binaires de 13,3 et 15,2 kbit/s. Généralement les codecs à bas débit exploitent les dépendances entre les trames. Le traitement indépendant des trames appliquées par iLBC offre une meilleure robustesse du codec, similaire à celle du codec G.711 avec le masquage de pertes de paquets (*Packet Loss Concealment* PLC).

3.2.8 Speex

Speex [36] est un codec libre (*opensource*) ciblée pour la VoIP. Le codec utilise CELP comme technique de codage. Speex supporte un large intervalle de débits : de 3,95 à 24,6 kbit/s pour les signaux à bande étroite (*narrowband*). L'encodage est contrôlé par le paramètre de qualité qui varie de 0 à 10. Le mode de plus faible qualité 0 (correspondant à un débit de 2,15 kbit/s) est principalement utilisé pour le bruit de confort (*confort noise*).

3.2.9 Silk

Le codec Silk [37] est utilisé par l'application Skype. Le débit binaire pour les signaux à bande étroite (*narrowband*) peut être réglée entre 6 et 20 kbit/s. Le codec fournit également la transmission discontinue

DTX (*discontinuous transmission*), un générateur de bruit de confort CNG (*Comfort Noise Generator*) et les mécanismes de masquage de pertes de paquets (*Packet Loss Concealment* PLC).

3.3 Les paramètres qui influencent la qualité d'expérience (QoE) de la VoIP

Il y a beaucoup de causes qui influent la perception de la qualité de la voix. Les facteurs dépendent, entre autres, de la technologie du réseau, de l'application, etc. Les paramètres, influant sur la qualité de la voix, peuvent être classés en quatre catégories différentes :

- Paramètres de l'environnement : Les paramètres de l'environnement caractérisent l'environnement dans lequel l'utilisateur consomme les médias correspondants. Nous pouvons citer, en tant que paramètres de l'environnement, le niveau de bruit ambiant de la salle, les haut-parleurs/microphone utilisées, les capacités du décodeur/ordinateur ... Ce type de paramètres est difficile à mesurer et le plus souvent incontrôlable.

- Paramètres source : Par « paramètres source » nous faisons référence aux paramètres liés au signal source. Par exemple, les paramètres tels que le niveau sonore, le codec utilisé, les techniques d'amélioration de la qualité (FEC *Forward Error Correction*) ... ont un impact évident sur la qualité perçue. Le codec et sa configuration jouent aussi un rôle important sur la qualité perçue. La qualité peut facilement varier d'un codec à un autre.

- Paramètres réseau : Les paramètres du réseau sont généralement les mêmes que les mesures de la qualité de service (QoS). Nous pouvons citer, comme exemples de paramètres de réseau, le taux de perte de paquets du réseau, la taille moyenne de pertes consécutives, le délai de bout-en-bout et/ou la gigue, ... L'impact des paramètres du réseau dépend de l'application utilisée. Par exemple, le taux de perte de paquets est un paramètre réseau important lorsqu'une distribution en temps réel est nécessaire. Le retard et la gigue ont aussi un impact important lors de l'utilisation des applications interactives. Nous pouvons noter aussi que les paramètres réseau sont généralement faciles à mesurer.

- Paramètres du récepteur : Les paramètres du récepteur portent sur les facteurs du côté de l'utilisateur final qui peuvent influer sur la qualité perçue. Par exemple, il existe un ensemble de procédures qui améliorent la qualité perçue par les utilisateurs finaux, tels que le masquage de perte de paquets et des améliorations du contrôle de congestion.

3.3.1 Paramètres réseau

La voix sur IP est un service de téléphonie utilisant le réseau Internet comme support au transport de la voix. Les échantillons de voix sont compressés à l'aide d'un codeur de parole et encapsulés dans des paquets de niveau transport. Pour la voix sur IP, le réseau servant de support au transport est le réseau Internet. Celui-ci fonctionne sur le modèle "*Best-Effort*". Chaque paquet est transporté du mieux possible d'un bout à l'autre du réseau.

Le protocole de transport généralement utilisé dans le cas de la voix sur IP (le cas de transmission de flux temps-réel plus généralement) est UDP (*User Datagram Protocol*). UDP est un protocole de communication en mode non-connecté. Dans le cas d'UDP, les paquets sont envoyés mais, contrairement au protocole TCP, chaque paquet est indépendant des autres et aucun mécanisme de contrôle (QoS) n'est mis en place pour assurer la bonne réception des paquets.

Le paquet transport est ensuite lui-même encapsulé dans un paquet IP (couche réseau) pour être transmis sur le réseau Internet. Lors de la transmission des paquets sur le réseau Internet, les paquets passent par divers réseaux de routeur et d'équipements, et peuvent ne pas emprunter le même trajet. De ce fait, les temps de parcours ne sont pas identiques pour tous les paquets et l'ordre d'arrivée des paquets diffère de leur ordre d'émission. Il est alors nécessaire avant de décompresser les paquets de les remettre dans l'ordre. C'est le rôle du tampon de compensation de gigue.

En plus du désordre des paquets, des phénomènes de congestion font que des équipements réseaux (routeurs) peuvent supprimer des paquets pour alléger la charge. On parle alors de « pertes de paquets ».

Dans les paragraphes suivants, nous décrierons les paramètres réseau les plus importants, qui ont un impact significatif sur la qualité de la voix perçu par l'auditeur.

3.3.1.1 Délai (*Delay*)

Lors des communications de données, il n'a pas vraiment d'importance lors d'un retard entre l'envoi d'un paquet et son arrivée. Cependant, avec la communication vocale, le retard global est extrêmement important. Le temps qui s'écoule entre une personne, qui dit quelque chose, et une autre qui écoute ce qui a été dit, doit être aussi faible que possible. Les grandes valeurs de retard résultent en une perte de l'interactivité. Dans le cas des communications en temps réel, du point de vue applicatif, le délai est généré en raison de la conversion analogique/numérique, de compression et de la décompression du signal, l'encapsulation en paquet, les files d'attente de l'interface réseau, la propagation et le temps d'attente dans le réseau, et le temps nécessaire pour la compensation de gigue (*dejittering*) du flux au niveau du récepteur.

Plusieurs études ont été menées afin de déterminer l'impact du délai sur les communications bidirectionnelles, en particulier pour les applications de téléphonie. Dans le standard ITU G.114, des seuils de retard ont été définis pour garantir une qualité acceptable de la conversation en termes d'interactivité. Bien sûr, les effets de retard sur la qualité dépendent en grande partie du type de l'application et les attentes de l'utilisateur.

Valeurs du délai (en ms)	Qualité de la conversation
< 150	Acceptable
150 – 400	Dégradation notable de l'interactivité
> 400	Inacceptable pour un usage général

Tableau 3-1. Influence d'un retard sur les communications interactives selon ITU G.114

Le Tableau 3-1 montre, selon l'ITU G.114 [38], l'influence du délai sur la qualité d'une conversation (en sens unique : *one-way delay*). En général, un délai inférieur à 150 ms permet de tenir une conversation satisfaisante. Les études montrent également qu'un délai de 150 à 400 ms est tolérable pour des courtes

durées d'une communication. Lorsque le retard dépasse 400 ms, une conversation téléphonique normale devient très difficile à tenir.

3.3.1.2 Gigue (*Jitter*)

Le temps nécessaire pour un paquet de passer d'un hôte à un autre sur Internet n'est généralement pas constant. Cette variation du délai est appelée gigue. Contrairement au délai, la gigue affecte d'une façon remarquable, à la fois les flux interactifs et celle à sens unique. Les flux multimédias sont généralement consommés à un rythme soutenu par le récepteur. Or, dans le cas où le paquet qui doit être lu n'est pas encore arrivé au récepteur, le système, dans l'obligation de transmettre les échantillons sonores, se comporte comme si le paquet était perdu.

Ce problème est généralement résolu en utilisant un tampon de compensation de gigue (*dejittering*), qui permet au client d'absorber les variations de retard sans perdre de paquet. Cependant, l'attente des paquets dans le tampon impliquerait une augmentation du délai de bout-en-bout, qui, dans certains cas (en particulier dans les applications interactives) peut conduire à une dégradation supplémentaire de la qualité perçue, comme expliqué dans le paragraphe précédent.

Par conséquent, en présence de gigue, il y a généralement un compromis à faire entre les paquets « perdus » (ceux qui arrivent trop tard pour être joué) et des retards. Beaucoup de recherches ont été menée dans ce domaine afin de d'évaluer l'impact du délai sur la qualité perçue, et de fournir une taille optimale du tampon, qui maximise la qualité du flux (par exemple : [39] [40]).

3.3.1.3 Perte de paquets

La perte de paquets est une source majeure de la distorsion de la parole dans la Voix sur IP. Les pertes de paquet peuvent être dues à plusieurs raisons : des problèmes de routage, les erreurs de transmission et la congestion du réseau. Les pertes de paquet au niveau du réseau sont normalement causées par la congestion (débordement de tampon des routeurs), l'instabilité de routage tel que les changements de route et les liens avec perte tels que les modems téléphoniques et les liens sans fil. La congestion est la cause la plus fréquente de perte de paquet [41].

Pour résumer, les causes majeures de pertes de paquets sont :

- le désordre à l'arrivée des paquets (*jitter*) ;
- la congestion dans les nœuds du réseau (routeurs) ;
- les erreurs de transmission ;
- le retard à l'arrivée du paquet.

Une étape importante dans le cas de perte de paquet est d'analyser et caractériser le processus de perte de bout-en-bout, afin de mieux comprendre comment il affecterait la qualité des applications. La caractérisation des processus de perte de paquet a fait l'objet d'un grand nombre de travaux. Plusieurs modèles ont été proposés pour la modélisation des caractéristiques des pertes de paquets [42][43][44]. Les modèles varient du très simple (des pertes indépendantes ou des fractions de perte de taille fixe), aux plus complexes (les chaînes de Markov). L'un des plus couramment utilisé, dans le cas de la VoIP, est une version simplifiée du modèle de *Gilbert* [45].

La plupart des recherches sur les communications VoIP utilisent le modèle *Gilbert* pour représenter les caractéristiques des pertes de paquets [46][47]. Dans le modèle de *Gilbert*, le canal présente deux états, l'un dans lequel la transmission est parfaite (état G pour *Good*), et un autre dans lequel une erreur se produit (état B pour *Bad*), comme le montre la Figure 3.2.

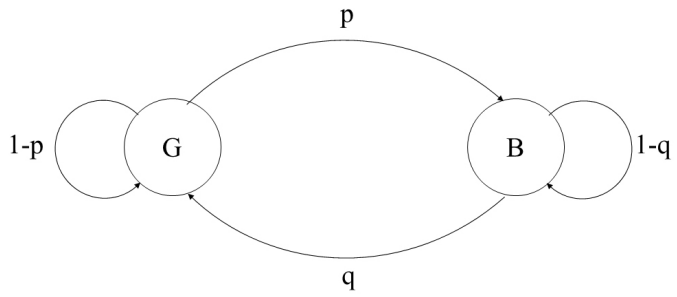


Figure 3.2. Modèle de Gilbert – Chaîne de Markov à 2-états

Nous définissons une variable aléatoire X comme suit : $X = 0$ (état G) pour un paquet reçu (sans perte) et $X = 1$ (état B) pour un paquet perdu. p est la probabilité qu'un paquet sera perdu dans le cas où le paquet précédent a été bien reçu. q est la probabilité qu'un paquet sera bien reçu dans le cas où le paquet précédent a été perdu.

Soit π_0 et π_1 les probabilités d'état pour l'état 0 (G) et 1 (B) :

$$\begin{aligned} \pi_0 &= P(X = 0) \\ \pi_1 &= P(X = 1) \end{aligned} \qquad \textbf{(3.1)}$$

La procédure pour calculer π_0 et π_1 est la suivante. A l'état d'équilibre, nous avons :

$$\begin{cases} \pi_0 = (1-p) \cdot \pi_0 + q \cdot \pi_1 \\ \pi_0 + \pi_1 = 1 \end{cases} \qquad \textbf{(3.2)}$$

Ainsi π_1, la probabilité de perte inconditionnelle, peut être calculée comme suit :

$$\begin{cases} \pi_0 = \dfrac{q}{p+q} \\ \pi_1 = \dfrac{p}{p+q} \end{cases} \qquad \textbf{(3.3)}$$

Le modèle Gilbert implique une répartition géométrique de la probabilité du nombre de paquets consécutifs perdus k, qui est la probabilité de perte consécutive ayant pour longueur k, p_k, qui est exprimée par :

$$p_k = P(LBS = k) = (1-q)^{k-1} \cdot q \quad , \quad k \geq 1 \qquad \textbf{(3.4)}$$

LBS (*Loss Burst Size*) est définie comme une variable qui décrit la distribution des longueurs des pertes consécutives de paquet par rapport aux événements de perte de paquets.

En se basant sur l'équation (3.4), la longueur moyenne des pertes consécutives *MLBS* peut être calculée comme suit :

$$MLBS = \sum_{k=1}^{\infty} k \cdot p_k = \sum_{k=1}^{\infty} k \cdot (1-q)^{k-1} \cdot q = \frac{1}{q} \quad (3.5)$$

Maintenant, notons par *PLR* le taux de perte de paquets (*Packet Loss Rate*), qui est la probabilité (inconditionnelle) de perdre un paquet :

$$\begin{aligned} \text{PLR} &= \pi_0 \cdot p + \pi_1 \cdot (1-q) \\ &= \frac{q}{p+q} \cdot p + \frac{p}{p+q}(1-q) \\ \text{PLR} &= \frac{p}{p+q} \\ &= \pi_1 \end{aligned} \quad (3.6)$$

A partir de l'équation (3.6), nous déduisons que :

$$p = q \cdot \frac{PLR}{1-PLR} = \frac{1}{MLBS} \cdot \frac{PLR}{1-PLR} \quad (3.7)$$

Maintenant que nous avons l'expression de *p* et *q* en fonction des paramètres *PLR* et *MLBS*, nous calibrons le modèle de Gilbert de la façon suivante :

$$\begin{aligned} q &= \frac{1}{MLBS} \\ p &= \frac{1}{MLBS} \cdot \frac{PLR}{1-PLR} \end{aligned} \quad (3.8)$$

3.3.2 Paramètres sources

3.3.2.1 Encodeurs

Lors de la transmission de flux multimédia, il est souvent souhaitable d'éliminer une partie des informations redondantes qui se trouvent dans les médias afin d'adapter le débit binaire du flux à la bande passante disponible. Le plus souvent, en raison des contraintes de bande passante, nous avons besoin de trouver une façon d'envoyer seulement les données nécessaires pour la reproduction du flux au niveau du récepteur.

L'utilisation de codecs avec perte permet d'atteindre des taux de compression plus élevés en sacrifiant certains « détails » du flux. Un exemple d'encodeur avec perte, est le MPEG Layer-3 (MP3) pour l'audio. Les codecs sans perte fournissent la même qualité que celle du signal d'origine, mais leurs taux de compression sont généralement trop faibles pour être utilisés dans les réseaux numériques.

Pour l'audio, nous avons trois grandes classes de techniques de codage, à savoir :

- **Encodage sous forme d'onde** (*waveform encoding*), qui maintient le signal codé très proche de celui d'origine. Une technique très simple qui utilise un codage de forme d'onde est *Pulse Code Modulation* (PCM), qui fait simplement l'échantillonnage et la quantification du signal. PCM est

une technique très simple, qui peut donner de très bonne qualité, au prix d'un débit assez élevé. Deux célèbres exemples de PCM sont les codecs *A-law* et *μ-law* (également connu sous le nom de l'ITU-T G.711). Des formes plus avancées de codage sous forme d'onde peuvent être trouvée dans les cas *Differential Pulse Code Modulation* (DPCM) et *Adaptive Differential Pulse Code Modulation* (ADPCM). Ces techniques exploitent la corrélation entre les échantillons de parole afin de prédire les futures formes d'onde à partir de celles du passé, et de transmettre l'erreur entre l'échantillon prédit et l'échantillon réel. La version adaptative de ce codec permet d'adapter le prédicteur pour correspondre aux caractéristiques de la parole étant actuellement codée.

- **Encodage de source**, qui permet d'atteindre des taux de compression très élevés en utilisant des modèles de parole en vue de réduire les données nécessaires pour la reconstruction du signal. L'idée principale est de déterminer les paramètres du modèle qui donnent le meilleur ajustement au signal d'origine et de les envoyer sur le réseau au lieu du signal lui-même. Du côté de la réception, les paramètres sont envoyés au modèle, et la parole est recréée et jouée. Ce type de codecs est également connu sous le nom « les vocodeurs » (*vocoder*). Les vocodeurs peuvent produire des résultats intelligibles à des débits faibles (2,4 kbit/s). La voix résultante ne dispose généralement pas d'une grande qualité, car elle est synthétisée, et perd donc des caractéristiques importantes. Un exemple bien connu de cette technique est l'algorithme de codage prédictif linéaire (*Linear Predictive Coding* LPC).

- **Encodage hybride** (*hybrid coding*) qui tente d'exploiter les avantages des deux techniques précédemment citées. Les encodeurs hybrides utilisent un modèle comme le cas des codeurs sources, mais ils essaient de mieux reconstituer les échantillons originaux en utilisant un signal d'excitation qui permet une meilleure adéquation entre l'échantillon reconstitué et celui d'origine. Le signal d'excitation est envoyé en même temps que les paramètres du modèle. Nous citons comme exemple de codeur hybride : *Residual Excited Linear Prediction* RELP, *Code-Excited Linear Prediction* CELP, *Multipulse and Regular Pulse Excited.* Plus de détails sur ces codeurs peuvent être trouvés dans [48]. Speex est un des codecs libre qui utilise la technique CELP.

3.3.2.2 Annulation d'écho

L' « écho » est un artefact bien connu des communications vocales, qui consiste en une ou plusieurs répétitions distinctes, retardées et atténuées du discours de l'interlocuteur. L'artefact Echo se produit généralement lorsque le son en provenance des haut-parleurs est renvoyé vers le microphone. Un autre scénario dans lequel l'écho peut se produire est un appel hybride où l'un des points finaux est un téléphone ordinaire, et l'autre est un périphérique VoIP.

Si le délai entre la parole et son écho est suffisamment petit, il peut passer inaperçue. Cependant, si le retard dépasse de quelques millisecondes, comme c'est souvent le cas, un suppresseur d'écho (*echo canceller*) doit être mis en place pour maintenir une qualité acceptable. À l'heure actuelle, la plupart des applications mettent en œuvre une certaine forme de suppression d'écho, afin d'éviter ce problème.

3.3.2.3 Détection et suppression de silence

Une communication voix se compose généralement d'une période de parole (*talk-spurts*) et d'une ou des périodes de silence. Comme aucune information utile n'est contenue dans les périodes de silence, alors ils peuvent être supprimés à partir du flux d'origine, ce qui réduit la consommation de bande passante.

La suppression de silence est effectuée par la détection de l'absence de parole par le biais d'un mécanisme de traitement de la parole appelé détection d'activité vocale (*Voice Activity Detection* VAD) qui fixe un seuil de détection de parole correspondant et génère dynamiquement un bruit de fond. Cette technique est également connue comme la détection d'activité vocale (*Speech Activity Detection* SAD).

3.3.2.4 Correction d'erreur : *Forward Error Correction* FEC

La correction d'erreurs *Forward Error Correction* (FEC) est basée sur l'ajout d'une redondance dans le flux, de manière à être en mesure de reconstituer au moins une partie des informations manquantes en raison de pertes de paquets dans le réseau. Il existe deux types de FEC pour les flux multimédias, à savoir :

- **FEC indépendant du média**, qui utilisent des techniques algébriques pour produire des paquets supplémentaires qui contiennent suffisamment d'informations pour reconstruire les paquets perdus (des exemples de ce type de FEC sont *parity coding*, *Reed–Solomon codes*, ou *complex XOR schemes*).
- **FEC dépendant du média**, qui bénéficient de la connaissance de la structure du flux et des propriétés de codage afin de fournir une protection efficace.

Les données supplémentaires introduites par le FEC ne sont généralement pas très significatives et leurs avantages sont très remarquables. Plus de détails sur les techniques du FEC peuvent être trouvés dans [49].

3.3.2.5 Entrelacement : *Interleaving*

L'entrelacement permet de disperser les effets de la perte de paquets sur le flux, afin de les rendre moins apparentes. L'idée est d'arranger les échantillons de telle manière que les échantillons consécutifs se retrouvent dans différents paquets. Ainsi, si un paquet est perdu, au lieu d'avoir un grand vide dans le flux, nous obtenons plusieurs interruptions plus petites. Si les interruptions sont assez petites, alors, elles ne seront pas faciles à détecter.

3.3.3 Paramètres réception

Outre les techniques mentionnées ci-dessus, qui sont utilisées au niveau de l'émetteur, plusieurs méthodes existent, pour l'amélioration de la qualité perçue, et sont mis en place au niveau du récepteur. Cette section fournit un bref aperçu des techniques les plus couramment utilisées.

3.3.3.1 L'utilisation des tampons de compensation de la gigue : *Buffering*

Les tampons sont utilisés au niveau du récepteur pour diminuer l'impact de la gigue (*jitter*) sur la qualité. Dans la plupart des applications, les paquets arrivent au décodeur à un débit constant. Toutefois, un des problèmes avec les réseaux de paquets comme Internet, c'est que le délai de bout-en-bout est assez variable. Par conséquent, si aucun tampon n'est présent au niveau du récepteur, les paquets peuvent

arriver plus tard que prévu, ce qui oblige à arrêter la lecture du flux, ou plus tôt que prévu, dans ce cas ils doivent être abandonnées.

Un tampon pour la compensation de la gigue (*dejittering*) aide à résoudre ce problème (ou du moins réduire son influence sur la qualité perçue), en éliminant les variations de délai entre les paquets. Bien sûr, le coût de cette amélioration est une augmentation du retard, car les paquets doivent rester au niveau de la mémoire tampon pendant un certain temps avant d'être décodées. La taille des tampons peut être ajustée dynamiquement, afin de minimiser le retard.

3.3.3.2 Masquage des pertes : *Loss Concealment*

On retrouve trois principaux types de techniques de dissimulation de pertes, à savoir :

- L'omission : Les paquets perdus sont remplacés par des trames de silence. Cette méthode bien que possédant un coût calculatoire nul produit un résultat de qualité médiocre : dès que la taille du trou dépasse les 20ms, l'auditeur perçoit une dégradation.
- La répétition : Les éléments de signal manquant sont remplacés par les paquets précédents. Il s'agit donc de répéter le ou les derniers paquets reçus. Ce type de masquage rend imperceptible des trous de petite taille.
- Les modèles de parole : De coût calculatoire plus important, ces méthodes extrapolent ou interpolent les paramètres d'un modèle de parole du signal reçu sur le signal manquant.

3.4 Evaluation de la qualité vocale

La mesure de la qualité de la parole peut être effectuée en utilisant soit des méthodes subjectives ou objectives comme le montre la Figure 3.3. La note moyenne d'opinion MOS est la mesure subjective de la qualité vocale la plus largement utilisée et est recommandée par l'ITU [38]. Nous rappelons que la valeur MOS est obtenue en moyennant les scores donnés par les observateurs pour évaluer la qualité vocale.

Comme le cas d'évaluation de la qualité vidéo, le problème inévitable des méthodes de mesure subjective telle que le MOS, consiste dans le fait qu'elles sont coûteuses, manquent d'automatisation et ne peuvent pas être utilisées pour la surveillance à grande échelle de la qualité de la voix dans une infrastructure réseau. Ces inconvénients ont rendu très attractives les méthodes objectives pour l'évaluation de la qualité afin de répondre à la demande pour la mesure de qualité de la voix dans les réseaux de communication.

La mesure objective de la qualité de la voix, dans les réseaux de communication modernes, peut être intrusive ou non intrusive. Les méthodes intrusives analysent les signaux vocaux transmis et reçus. Ces dernières comparent le signal vocal de référence avec le signal correspondant déformé. Les méthodes non intrusives permettent une estimation de la qualité vocale perçue en exploitant des informations extraites du côté du récepteur. Les méthodes non intrusives utilisent uniquement le signal de parole dégradé (reçu) pour estimer la qualité vocale correspondante.

Les méthodes intrusives sont plus précises que les non intrusives, mais elles ne sont pas adaptées pour le suivi en temps réel du trafic à cause de la nécessité de disposer des données de référence. Une

méthode typique intrusive est basée sur la dernière norme ITU P.862, *Perceptual Evaluation of Speech Quality* (PESQ) [51].

Les méthodes non intrusives sont plus appropriées pour la surveillance du trafic en temps réel vu qu'elles n'ont pas besoin du signal de référence. Il y a deux catégories de méthodes non intrusives : celles qui sont basées sur le signal dégradé reçu (*signal-based method*) et celles qui sont basées sur des paramètres du réseau ou de la voix (*parameter-based method*). Un exemple de méthode non-intrusive basée sur le signal est le modèle d'appareil vocal : *vocal tract model* [52], qui vise à prédire la qualité vocale en analysant directement le signal de parole en cours d'écoute (un signal dégradé) sans le signal de référence. L'exemple le plus connu et le plus largement utilisé des méthodes non intrusives paramétrique, est le ITU-T *E-model* [53]. Ce modèle peut estimer le score MOS d'une conversation directement à partir des paramètres du réseau IP et/ou du terminal.

Nous présentons dans cette section les différentes méthodes de mesure subjectives, ainsi que les méthodes objectives intrusives et non intrusives dédiés à la voix.

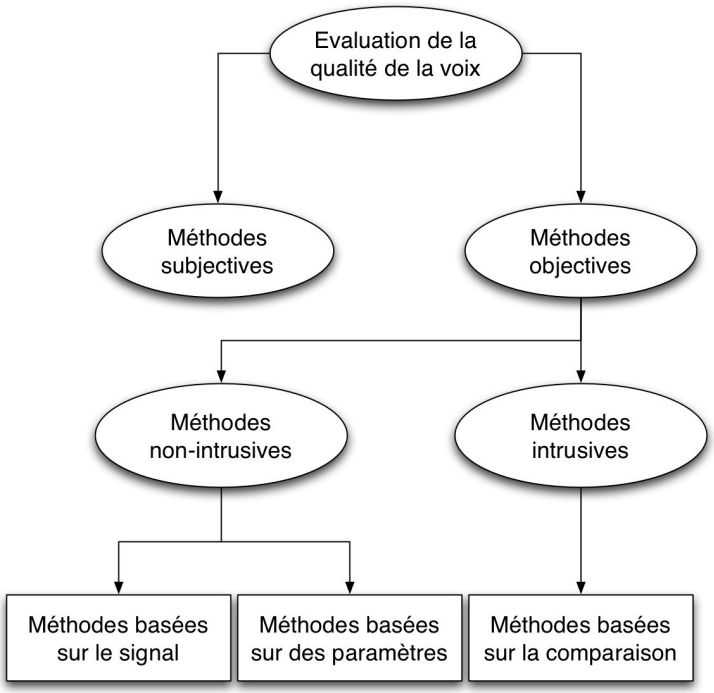


Figure 3.3. Classification des méthodes d'évaluation de la qualité de la parole

3.4.1 Mesure subjective de la qualité vocale

Comme pour la vidéo, les méthodes subjectives sont cruciales pour l'analyse comparative des méthodes objectives. L'ITU P.800 [50] décrit plusieurs méthodes et procédures pour la conduite des évaluations subjectives de la qualité de transmission. La méthode la plus couramment utilisée est l'évaluation par catégories absolues (*Absolute Category Rating* ACR) qui donne la note moyenne d'opinion (MOS). La méthode d'évaluation par catégories de dégradation (*Degradation Category Rating* DCR) est également utilisée dans certaines occasions. Cette méthode donne la dégradation de la note moyenne d'opinion : *Degradation Mean Opinion Score* (DMOS).

L'évaluation subjective de MOS est habituellement effectuée dans des conditions contrôlées en laboratoire (par exemple, dans une salle insonorisée).

3.4.1.1 Evaluation par catégories absolues : *Absolute Category Rating* (ACR)

Les tests ACR sont le plus couramment utilisés pour évaluer la qualité intégrale de la parole (ITU-T Rec. P.800 [50]). Dans ce genre de test, un groupe d'auditeurs évalue une série de fichiers audio (voix) à l'aide d'une échelle d'appréciation à cinq niveaux de valeur (Tableau 3-2), sans avoir à écouter la séquence originale. La quantité évaluée d'après les notes (note moyenne d'appréciation de qualité d'écoute ou simplement note moyenne d'appréciation) est représentée par le symbole MOS.

Qualité de la parole	Score
Mauvaise	1
Médiocre	2
Passable	3
Bonne	4
Excellente	5

Tableau 3-2. Echelle d'appréciation de la qualité d'écoute

3.4.1.2 Evaluation par catégories de dégradation : *Degradation Category Rating* (DCR)

Lorsque des échantillons de parole de bonne qualité sont évalués, l'ACR a tendance à être insensible aux petites dégradations de qualité. La procédure d'évaluation par catégories de dégradation DCR, qui repose en particulier sur une échelle de perturbation et sur une référence de qualité élevée avant chaque configuration à évaluer, semble convenir pour évaluer la parole de bonne qualité. Les sujets notent le niveau de dégradation et de gêne en comparant avec le signal vocal d'origine (de référence). Les échelles de notation (les niveaux de dégradation) sont présentées dans le Tableau 3-3.

La quantité de dégradation évaluée d'après les notes (notes d'appréciation moyenne de la dégradation) est représentée par le symbole DMOS.

Niveau de dégradation	Score
Dégradation inaudible	5
Dégradation audible mais pas gênante	4
Dégradation peu gênante	3
Dégradation gênante	2
Dégradation très gênante	1

Tableau 3-3. Echelle de notation pour les tests DCR

Afin de normaliser les tests subjectifs, ITU P.800 définit des conditions détaillées des essais subjectifs tels que les caractéristiques des matériaux d'essai et l'environnement de test et d'essai. Les tests subjectifs sont normalement effectués dans une zone réglementée, à double paroi, chambre insonorisée.

Les principaux inconvénients de l'évaluation subjective (auditive) sont : (i) la nécessité des méthodes spécialisées, (ii) le besoin de personnel ayant une expertise spécifique, (iii) la nécessité de beaucoup de temps, (iv) l'impossibilité de les utiliser pour une mesure en temps réel. Ces limitations ont motivé le développement de méthodes objectives (appelée aussi instrumentales) qui mesurent la qualité vocale d'une manière systématique.

3.4.2 Mesure objective (instrumentale) intrusive de la qualité vocale

Afin de réduire le temps et les coûts nécessaires pour mesurer la qualité de la voix, des mesures objectives (appelées aussi instrumentales) ont été proposées. Les méthodes objectives sont des algorithmes ou des dispositifs, qui agissent comme un instrument de mesure, et génèrent une mesure quantitative à partir des échantillons de la voix et la faire correspondre à une prédiction de qualité. La qualité prédite doit être très proche de la valeur réelle de la qualité perçue de la voix. Habituellement, la précision d'une méthode instrumentale est déterminée par sa corrélation avec les scores MOS pour un grand ensemble de données obtenues dans une variété de systèmes et de conditions de transmission.

Les méthodes objectives intrusives d'évaluation de la qualité peuvent être classées en trois groupes, chacune ayant une certaine correspondance avec les trois techniques de codage mentionné dans la section 3.1.

3.4.2.1 Mesures dans le domaine temporel

Les mesures dans le domaine temporel sont surtout destinées à des systèmes de codage/transmission qui tentent de reproduire la forme d'onde d'origine (par exemple, dans le cas des techniques de codage de forme d'onde), et nécessitent une grande précision d'alignement temporel des flux afin de comparer les deux formes d'onde.

Les mesures dans le domaine temporel, les plus simples et les plus communes, pour l'évaluation de la qualité du signal sont *Signal-to-Noise-Ratio* (SNR) and *Segmental Signal-to-Noise-Ratio* (SSNR).

Soient *x(i)* la parole d'origine et *y(i)* la version déformée du signal, le SNR est défini par

$$SNR = 10\log_{10}\frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}\big(x(i)-y(i)\big)^2} \qquad \textbf{(3.9)}$$

avec *i* l'indice dans le temps couvrant la période de mesure.

SSNR est une variation du SNR qui fonctionne sur des segments courts (15 à 20 ms) de flux. Cela permet de faciliter l'alignement temporel et il fournit des résultats qui sont légèrement mieux (en termes de corrélation avec l'évaluation subjective) que ceux du SNR. Il est calculé comme suit :

$$SSNR = \frac{1}{N}\sum_{m=1}^{N} SNR_m \qquad \textbf{(3.10)}$$

Il s'agit d'une moyenne des valeurs SNR obtenues pour des trames isolées. Les trames étant un bloc d'échantillons.

Les mesures dans le domaine temporel sont faciles à mettre en œuvre, et peuvent être utiles pour détecter les distorsions introduites par un bruit additif ou le bruit généré par le codeur de forme d'onde. Cependant, ils montrent des limitations lorsqu'elles sont utilisées dans un contexte général surtout quand il y a des dégradations telles que le filtrage ou les distorsions de phase.

3.4.2.2 Mesures dans le domaine spectrales

Les mesures dans le domaine spectrales sont généralement appliquées à des segments courts, typiquement de l'ordre de 15 à 30 ms. Ces techniques sont moins sensibles aux problèmes d'alignement de temps que les mesures dans le domaine temporel, et elles fournissent en général de meilleurs résultats. Elles sont, cependant, très dépendantes du codec utilisé (ce qui limite leur généralité) et ne conviennent pas pour une utilisation dans les réseaux, qui peuvent modifier certaines caractéristiques des flux. Ce type de mesure est bien adapté pour l'évaluation des distorsions causées lors du codage. Les méthodes les plus connues dans le domaine fréquentiel sont *Itakura-Saito* (IS) et *Log-Area-Ratio* (LAR) [54].

3.4.2.3 Mesures dans le domaine perceptuel

Les mesures dans le domaine perceptuel sont basées sur une certaine forme de modèle psycho-acoustique. Elles sont généralement plus précises que les autres types de mesure. Le point faible de ces mesures est le fait qu'elles sont généralement optimisées pour un certain type de communication, et donc elles ne peuvent pas fournir des évaluations précises dans le cas d'autres types de données vocales.

Plusieurs méthodes perceptives pour l'estimation de la qualité vocale ont été développées : BSD [55], TOSQA [56], PAMS [57], PSQM [58], PSQM+ [59], and PESQ [51], POLQA [60] ...

3.4.2.3.1 Perceptual Speech Quality Measure (PSQM)

PSQM, développé par *PTT research* en 1994, est une version modifiée de *Perceptual Audio Quality Measure* (PAQM), qui est une mesure objective de la qualité de la voix. C'était la première mesure intrusive pour la qualité de parole à bande étroite à être normalisée par l'ITU-T dans la recommandation P.861[58]. PSQM est un algorithme d'évaluation perceptuelle de la qualité de la parole, conçu pour évaluer la performance des codecs vocaux et des distorsions rencontrées dans les réseaux. Il utilise les résultats de calcul psycho-acoustiques de l'intensité vocale pour transformer la parole dans le domaine perceptuel correspondant. Le modèle PSQM donne des estimations fiables dans le cas des codecs vocaux à faible débit binaire, mais il peut ne pas être suffisamment robuste pour être appliquée à un plus grand éventail de distorsions. Une version améliorée du modèle PSQM, appelé PSQM+ [61], a été développée. PSQM+ prédit les distorsions dues aux erreurs de canal de transmission, telle que la perte de paquets.

3.4.2.3.2 Perceptual Analysis Measurement System (PAMS)

British Telecom (BT) a élaboré en 1998 le système d'analyse de mesure perceptuelle *Perceptual Analysis Measurement System* (PAMS) [62]. PAMS compare le signal original avec le signal dégradé en transformant les distorsions en paramètres et les faire correspondre à une mesure objective de la qualité. Le modèle PAMS fournit un algorithme efficace pour l'alignement temporel afin d'estimer avec précision le temps de retard introduit par le système de transmission. Ainsi, contrairement à d'autres modèles, le PAMS est capable d'aligner le signal original et dégradé, dans le cas d'un retard variable dans la séquence vocale entière.

3.4.2.3.3 Perceptual Evaluation of Speech Quality (PESQ)

Les mesures les plus efficaces, évalués par l'ITU dans les années 1990, ont été combinés en un modèle amélioré *Perceptual Evaluation of Speech Quality* (PESQ), qui a été accepté en tant que recommandation de l'ITU en 2001 [51].

PESQ est considéré comme une combinaison optimale de deux algorithmes PSQM et PAMS. PESQ a hérité les deux composants suivant : (i) l'algorithme d'alignement du modèle PAMS [63] et (ii) la transformation dans le domaine perceptuel du modèle PSQM99 [64]. L'alignement temporel consiste à calculer l'ensemble des retards entre le signal original et le signal déformé. La transformation perceptive est utilisée pour transformer, à la fois le signal d'origine et dégradé, à une représentation psychophysique dans le système auditif humain, en tenant compte de l'intensité et de la fréquence perceptuelle.

PESQ est un outil puissant pour mesurer une qualité vocale lorsqu'il est utilisé judicieusement. La mesure objective de la qualité de la voix par PESQ est fortement corrélée avec les mesures subjectives. Toutefois, il est connu que PESQ n'est pas destiné à être utilisé, ou génère des résultats erronés, pour la mesure de qualité de la voix conversationnelle (communication dans les deux sens) lors des tests subjectifs. Notons que le résultat de PESQ est un score qui varie entre -0,5 (une très mauvaise qualité) et 4,5 (une excellente qualité). Une version large bande de PESQ (WB-PESQ) a été définie dans la Recommandation ITU-T Rec. P.862.2 (2005) [65]. WB-PESQ utilise exclusivement des signaux de parole avec une fréquence d'échantillonnage de 16 kHz.

3.4.2.3.4 Perceptual Objective Listening Quality Analysis (POLQA)

POLQA [60] est la technologie de la prochaine génération de mesure de la qualité de la voix pour la téléphonie fixe, mobile et IP. POLQA est considéré comme le successeur de PESQ. POLQA a été normalisé par l'Union Internationale des télécommunications ITU-T en tant que nouvelle Recommandation P.863, et peut être appliquée à l'analyse de la qualité de la voix dans le contexte HD Voice, 3G et réseaux 4G/LTE. Il s'agit d'un modèle intrusif pour la mesure de la qualité vocal, convenable pour les interfaces électro-acoustiques et les connexions à bande étroite (*NarrowBand*) et à bande super-large (*S-WideBand*). POLQA utilise un modèle psycho-acoustique pour émuler la perception humaine.

3.4.3 Mesure objective non-intrusive de la qualité vocale

Contrairement aux méthodes intrusives, dans lesquelles un signal de référence est nécessaire pour l'évaluation de la qualité, les méthodes de mesure non intrusives de la qualité de la parole n'ont pas besoin du signal de référence et sont plus appropriées pour la surveillance du trafic en temps réel.

Il existe deux catégories de méthodes non intrusives pour la prévision de la qualité de parole. La première consiste à prédire la qualité de la parole directement à partir de divers valeurs de paramètres du réseau (par exemple, la perte de paquets, la gigue et le retard) et les paramètres non liés au réseau (par exemple, le codec, écho, la langue, etc.). Le but est d'établir la relation entre la qualité vocale perçue et les paramètres associés du réseau ou hors réseau. Les méthodes typiques sont le E-model et les réseaux de neurones artificiels (*Artificial Neural Network* ANN).

3.4.3.1 E-model

L'ETSI a développé un modèle de calcul de la qualité de transport de la voix de bout en bout, de la bouche de l'émetteur à l'oreille du récepteur, connu sous le nom de E-model (référence ETSI ETR 250 [66]). Ce modèle a été standardisé par l'ITU sous la référence G.107 [53]. Le principe de l'E-model consiste à calculer une grandeur unique *R*, qui reflète l'état de la conversation, en fonction des paramètres et caractéristiques des équipements ainsi que celles du canal de transmission du signal.

Le E-model a été développé à l'origine pour la planification du réseau, mais il est maintenant utilisé pour prédire la qualité de voix d'une façon non intrusive pour les applications VoIP.

La formule simplifiée du calcul de *R* est la suivante :

$$R = R_0 - I_s - I_d - I_{e-eff} + A \qquad \textbf{(3.11)}$$

Avec :

- R_0 : coefficient initial signal sur bruit. R_0 est la valeur que l'on obtiendrait si la transmission était parfaite
- I_s : coefficient de dommages simultanés avec l'émission de la voix
- I_d : coefficient de dommages dus au délai de transmission et de transport
- I_{e-eff} : coefficient de dommages de distorsion causés par les équipements
- A : coefficient de prise en compte de facteurs d'amélioration du réseau. Par exemple, les utilisateurs de services mobiles sont plus tolérants à l'égard d'un canal dégradé que les utilisateurs de réseaux câblés.

Le facteur de transmission de R estimé par E-model peut être converti en un MOS en utilisant l'équation (B-03) de l'annexe B de la recommandation ITU-T G.107 [53] :

$$MOS = \begin{cases} 1 & R < 0 \\ 1 + 0{,}035 \cdot R + 7 \cdot 10^{-6} \cdot R(R-60)(100-R) & 0 < R < 100 \\ 4{,}5 & R > 100 \end{cases} \qquad \textbf{(3.12)}$$

E-model est un modèle attractif pour la prédiction non-intrusive de la qualité de voix, mais il a un certain nombre de limitations. Par exemple, il est basé sur un ensemble complexe de formules fixes, empiriques et applicables à un nombre limité de codecs et de conditions de réseau (parce que les tests subjectifs sont nécessaires pour obtenir les paramètres du modèle). En outre, E-model est un modèle statique qui ne peut pas s'adapter à l'environnement dynamique des réseaux IP.

3.4.3.2 IQX Hypothesis

IQX Hypothesis vise à décrire la qualité perceptuelle ou plus connue sous le nom de *Quality of Experience* en fonction d'un facteur unique qui détermine la qualité de service du réseau (par exemple taux de perte de paquets). Cette hypothèse exprime la QoE comme une fonction exponentielle de la dégradation de qualité de service :

$$QoE = \alpha \cdot e^{-\beta \cdot QoS} + \gamma \qquad \textbf{(3.13)}$$

Pour valider le modèle mathématique, des échantillons de parole, codées avec le codec iLBC, ont été transmis sur un émulateur de réseau avec des taux de perte de paquets variables, et avec leurs sources correspondantes ont servi de base pour l'application de l'algorithme PESQ. Les résultats ont prouvé que, dans le cas du codec iLBC, la relation entre la QoE et le facteur de dépréciation (perte de paquets) est comme suit :

$$QoE = 3{,}010 * \exp(-4{,}473 * p_{loss}) + 1{,}065 \quad \textbf{(3.14)}$$

3.4.3.3 Les réseaux de neurones artificiels

Contrairement à la méthode E-model qui est un modèle de calcul mathématique et qui est statique, les modèles basés sur les réseaux de neurones artificiels (*Artificial Neural Networks* ANN) peuvent s'adapter à l'environnement dynamique des réseaux IP, en raison de sa capacité d'apprentissage. Un modèle ANN peut être construit par l'apprentissage des relations non linéaires entre la qualité vocale perçue (par exemple, note MOS) et une variété de paramètres réseau ou liés à la parole.

Les réseaux de neurones artificiels seront plus détaillés dans le chapitre 4.

Conclusion

Nous avons présenté dans ce chapitre le principe de la Voix sur IP et les codecs les plus utilisés par les logiciels de communication VoIP. Nous avons cité aussi les méthodes d'évaluation (objectives et subjectives) de la voix.

Bon nombre de mesures objectives, décrites dans ce chapitre, ont été initialement développées pour mesurer les dégradations subies par la parole codée. Par conséquent, ils ne prennent généralement pas en compte les facteurs du réseau, et leurs performances sont affectées quand elles sont utilisées pour évaluer les flux VoIP. Par ailleurs, à l'exception du E-model et de l'IQX, l'ensemble de ces méthodes propose différentes façons de comparer le signal reçu avec le signal original. Cela empêche leur utilisation en temps réel, ce qui est essentiel pour les applications de contrôle de la qualité.

Nous décrivons dans le chapitre suivant le principe des réseaux de neurones artificiels et leurs applications pour l'estimation de la QoE.

Chapitre 4

4 La méthode PSQA

Pseudo-Subjective Quality Assessment [67] est une méthodologie pour la mesure non-paramétrique sans référence de la QoE. PSQA peut être considérée comme une méthode hybride entre les techniques d'évaluation subjective et objective. L'idée principale, illustrée dans la Figure 4.1, est d'avoir plusieurs échantillons déformés évalués subjectivement, puis d'utiliser les résultats de cette évaluation pour entrainer un réseau de neurones afin de capturer la relation entre (i) les paramètres qui provoquent la distorsion et (ii) la qualité perçue. La procédure consiste à choisir un ensemble de paramètres P (sélectionné a priori), qui peuvent avoir un impact sur la qualité perçue. Par exemple, nous pouvons choisir le codec utilisé, le taux de perte de paquets du réseau, la taille moyenne de perte en rafale, le délai de bout-en-bout et/ou la gigue, etc. Supposons que cet ensemble soit noté par $P = \{\pi_1, \pi_2, ..., \pi_n\}$. Une fois que les paramètres affectant la qualité sont définis, il est nécessaire de choisir un ensemble de valeurs représentatives pour chacun d'eux, avec un intervalle dans lequel les paramètres varient selon les conditions dans lesquelles nous souhaitons que le système fonctionne. Le nombre de valeurs à choisir pour chaque paramètre dépend de la taille de l'intervalle choisi et de la précision souhaitée. Par exemple, si nous considérons le taux de perte de paquets comme l'un des paramètres, et si nous nous attendons à ce que ces valeurs varient principalement de 0 à 5%, nous pourrons alors choisir les taux de perte suivants : 0%, 1%, 2%, 3%, 4% et 5%. Dans ce contexte, nous appelons une « configuration » l'ensemble de la forme $\Omega = \{v_1, v_2, ..., v_n\}$, où v_i est l'une des valeurs choisies pour le paramètre p_i.

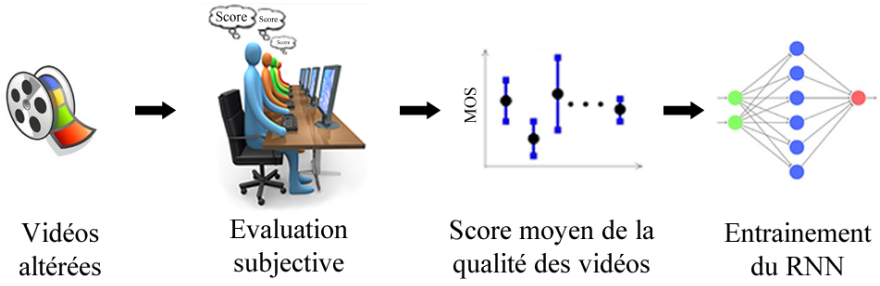


Figure 4.1. Les étapes d'entrainement du réseau de neurones

Le nombre total de configurations possibles est généralement très volumineux. Par conséquent, l'étape suivante consiste à sélectionner un sous-ensemble de configurations S à évaluer subjectivement. Cette sélection peut se faire au hasard, mais il est important de couvrir les points situés près des frontières de l'espace de configuration. Il n'est alors pas nécessaire d'utiliser une distribution uniforme. Il faut échantillonner plus de points dans les régions proches des configurations qui sont les plus susceptibles de se produire pendant l'utilisation normale ou ceux considérés comme les plus importants pour une raison quelconque. Une fois que les configurations ont été choisies, nous devons construire un ensemble d'« échantillons déformés », c'est-à-dire des échantillons résultant de la transmission des données d'origine

sur le réseau dans le cadre des différentes configurations choisies. Pour cela, nous utilisons un banc d'essai ou un simulateur de réseau, ou une combinaison des deux.

Nous devons maintenant sélectionner un ensemble d'échantillons de médias *M*, noté (∂_m) avec $m=1, .., M$, par exemple, *M* séquences courtes de vidéo (les normes de tests subjectives conseillent l'utilisation des séquences ayant une longueur de 10s). Nous notons l'ensemble des configurations *S* déjà échantillonnées par $\{\Omega_1, ..., \Omega_S\}$ où $\Omega_S = \{v_{s1}, ...,v_{sp}\}$, avec v_{sp} étant la valeur du paramètre π_p dans la configuration Ω_S. A partir de chaque échantillon ∂_i, nous construisons un ensemble d'échantillons $\{\partial_{i1}, ..., \partial_{iS}\}$ qui ont rencontrés des conditions variées lors de la transmission sur le réseau de la manière suivante : la séquence ∂_{iS} est la séquence qui est arrivée au niveau du récepteur lorsque l'expéditeur a envoyé la séquence ∂_i sur la plateforme de test où les paramètres *P* avaient les valeurs de la configuration Ω_S.

Quand les échantillons déformés sont générés, un test subjectif est effectué sur chaque séquence ∂_{iS} reçue. Les scores obtenus subissent un traitement statistique. Ce traitement statistique est conçu pour détecter et éliminer les mauvaises notes. Des notes sont considérées mauvaises quand elles ne sont pas statistiquement cohérentes avec la majorité. Après le traitement statistiques des notes, la séquence ∂_{iS} reçoit une note μ_{iS} (souvent, il s'agit d'un *Mean Opinion Score* : moyenne de tous les scores). L'idée est alors d'associer à chaque configuration, la valeur :

$$\mu_s = \frac{1}{M}\sum_{m=1}^{M} \mu_{ms} \qquad \textbf{(4.1)}$$

A ce stade, il existe un score de qualité associée à chaque configuration Ω_S. Nous choisissons maintenant S_1 configurations au hasard parmi les *S* configurations disponibles. Ces configurations, ainsi que leurs valeurs, constituent la « base de données d'entraînement ». Les autres configurations $S_2=S-S_1$ et leurs valeurs correspondantes constituent la « base de données de validation », réservée à la dernière (et critique) étape du processus. La phase suivante du processus consiste à entrainer le réseau de neurones pour apprendre la correspondance entre les configurations et les scores tels que définis par la base de données d'entraînement.

Supposons que les paramètres sélectionnés ont des valeurs (mises à l'échelle) dans [0, 1] et la même chose avec les scores de qualité. Une fois que le réseau de neurones a été entrainé, nous avons une fonction *f* de $[0,1]^p$ dans [0,1], qui permet de faire correspondre un score à chaque configuration possible. La dernière étape est la phase de validation : nous comparons la valeur donnée par *f* d'un point correspondant à chaque configuration de Ω_S de la base de données de validation, avec les scores de qualité μ_{iS} correspondant. Si les scores donnés par le réseau de neurones sont assez proches des scores subjectifs, alors l'entraînement est validé. Si la validation échoue, il faut revoir l'architecture choisie du réseau de neurones et les configurations choisies.

4.1 Les réseaux de neurones aléatoire (RNN)

Les réseaux de neurones aléatoires (*Random Neural Network* RNN), inspiré par le comportement des réseaux de neurones biologiques, étaient développés par E. *Gelenbe* [67]. Les RNN sont considérés comme une fusion entre les réseaux de neurones artificiels et les réseaux de files d'attente.

RNN est composé d'un ensemble de neurones interconnectés qui échange des signaux. Ces signaux peuvent être de deux types, excitateurs (positifs) ou inhibiteurs (négatifs) : excitateurs, ils vont faire augmenter le potentiel du neurone qui les perçoit; inhibiteurs, ils vont faire diminuer ce potentiel. Si ce potentiel est positif, le neurone pourrait alors lui-même émettre des signaux.

Un neurone aléatoire *i* est caractérisé par son état interne à valeur dans $\mathbb{N}$, appelé potentiel, à l'instant *t*, noté $k_i(t)$. La valeur de $k_i(t)$ dépend de trois facteurs :

- s'il reçoit un signal excitateur, alors son potentiel augmente d'une unité ;
- s'il reçoit un signal inhibiteur, alors son potentiel diminue d'une unité ;
- s'il émet un signal, son potentiel diminue d'une unité.

Nous appelons un neurone excité, celui qui a un potentiel strictement positif ($k_i(t) > 0$). Les neurones excités émettent aléatoirement des signaux aux autres neurones (ou à l'environnement), selon un processus de *Poisson* avec une fréquence r_i.

Pour une meilleure lecture et compréhension du texte, les symboles du modèle sont présentés dans le Tableau 4-1.

Notation	Définition
N	Nombre de neurone dans le réseau
$\boldsymbol{k_i(t)}$	Potentiel du neurone *i* à l'instant *t*
$\boldsymbol{\varrho_i(t)}$	Probabilité que le neurone *i* soit excité à l'instant *t*
$\boldsymbol{r_i}$	Taux d'envoi de signaux du neurone *i*
$\boldsymbol{d_i}$	Probabilité qu'un signal émis par le neurone *i* quitte le réseau
$\boldsymbol{\lambda_i^+}$	Taux d'arrivée extérieur de signaux positifs au neurone *i*
$\boldsymbol{\lambda_i^-}$	Taux d'arrivée extérieur de signaux négatifs au neurone *i*
$\boldsymbol{T_i^+}$	Taux d'arrivée de signaux positifs au neurone *i*
$\boldsymbol{T_i^-}$	Taux d'arrivée de signaux négatifs au neurone *i*
$\boldsymbol{p_{ij}^+}$	Probabilité que le neurone *j* reçoit un signal positif d'un neurone *i*
$\boldsymbol{p_{ij}^-}$	Probabilité que le neurone *j* reçoit un signal négatif d'un neurone *i*
$\boldsymbol{w_{ij}^+}$	Taux de signaux positifs envoyés par neurone *i* au neurone *j*
$\boldsymbol{w_{ij}^-}$	Taux de signaux négatifs envoyés par neurone *i* au neurone *j*

Tableau 4-1. Résumé des notations

La probabilité que le signal envoyé par le neurone *i* va au neurone *j* soit excitateur, est désigné par p_{ij}^{+} et par p_{ij}^{-} s'il est inhibiteur. Les signaux quittent le réseau avec une probabilité d_i. Donc, pour un réseau de *N* neurones, nous avons pour tout $i = 1,..,N$:

$$d_i + \sum_{j=1}^{N} \left(p_{ij}^{+} + p_{ij}^{-}\right) = 1 \qquad \textbf{(4.2)}$$

Par analogie avec les réseaux de neurones classiques, nous parlons de poids des connexions. Ce sont les poids qui déterminent le comportement du réseau. Il existe deux poids par connexion, notés w_{ij}^{+} et w_{ij}^{-}. Ainsi, lorsque le neurone *i* est excité, il envoi des signaux positifs et négatifs au neurone *j* avec des taux :

$$\begin{cases} w_{ij}^{+} = r_i p_{ij}^{+} \geq 0 \\ w_{ij}^{-} = r_i p_{ij}^{-} \geq 0 \end{cases} \qquad \textbf{(4.3)}$$

L'état du réseau est spécifié, à l'instant *t*, par le potentiel de ses neurones : $\vec{k}(t) = (k_1, \ldots, k_n)$, avec k_i le potentiel du neurone *i* à l'instant *t*. *Gelenbe* a prouvé que dans le cas d'un processus ergodique $\vec{k}(t)$ (réseau stable), nous avons la distribution de probabilité conjointe suivante :

$$\Pr\left(\vec{k}(t) = (k_1, \ldots, k_N)\right) = \prod_{i=1}^{N} (1 - \varrho_i)\varrho_i^{k_i} \qquad \textbf{(4.4)}$$

ϱ_i, un paramètre, étroitement lié à $k_i(t)$, est la probabilité d'excitation du neurone *i* :

$$\varrho_i = \lim_{t \to \infty} Pr[k_i(t) > 0] \leq 1 \qquad \textbf{(4.5)}$$

On peut alors définir les grandeurs caractéristiques du neurone :

- la fréquence d'arrivée des signaux excitateurs au neurone *i*, notée T_i^{+};
- la fréquence d'arrivée des signaux inhibiteurs au neurone *i*, notée T_i^{-};
- la fréquence d'émission des signaux par le neurone *i*, notée r_i.

Le potentiel du neurone *i* se calcule donc ainsi :

$$\forall\, i\,, \qquad \varrho_i = \frac{T_i^{+}}{r_i + T_i^{-}} \qquad \textbf{(4.6)}$$

Avec

$$\forall\, i\,, \qquad T_i^{+} = \lambda_i^{+} + \sum_{j=1}^{N} \varrho_j w_{ji}^{+} \qquad \textbf{(4.7)}$$

$$\forall\, i\,, \qquad T_i^{-} = \lambda_i^{-} + \sum_{j=1}^{N} \varrho_j w_{ji}^{-} \qquad \textbf{(4.8)}$$

Gelenbe a prouvé que ce système non-linéaire a une solution unique, et le réseau est stable si et seulement si ce système non-linéaire a une solution où pour tout $i \in \{1,..,N\}, \varrho_i < 1$. Plus de détails sur les réseaux de neurones aléatoires peuvent être trouvés dans [67][69][70].

La solution de ce système d'équation pourrait être trouvée en utilisant un algorithme de minimisation tels que l'algorithme de descente de gradient tel que proposé par *Gelenbe* [67], ou avec des techniques plus sophistiquées, comme la méthode de *Levenberg-Marquardt* [71][72] ou l'optimisation de quasi-Newton proposé par *Aristidis Likas* et *Andreas Stafylopatis* [73].

4.2 Comparaison des résultats de PSQA avec d'autres outils objectives pour l'évaluation de la qualité des vidéos MPEG-2

Dans cette section, nous allons comparer les résultats de PSQA avec d'autres outils d'évaluation de la qualité basés sur le modèle objectif. Ceci va nous permettre de mettre en avant l'avantage de PSQA par rapport à d'autre méthode et valider notre choix d'utiliser PSQA pour la suite. Nous nous intéressons à l'estimation de la qualité des vidéos encodées en MPEG-2. Nous présentons l'environnement de test dans la première partie, et nous présentons les résultats obtenus avec la méthode PSQA. Ces résultats seront comparés à ceux obtenus avec deux méthodes objectives sans référence (Demokritos, Bsoft) et deux méthodes objectives avec référence (SSIM et PSNR).

4.2.1 Environnement de test

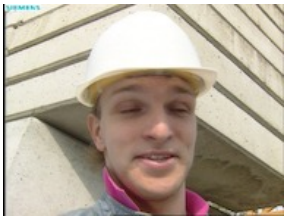
Notre étude s'appuie sur une approche expérimentale. A cette fin, nous avons mis en place une plate-forme permettant l'émulation d'une transmission d'un flux vidéo MPEG-2 avec pertes. Nous avons utilisé des séquences vidéo courtes avec des quantités d'information spatiales et temporelles différentes. Notre choix des séquences vidéo pour le test s'est porté sur quatre séquences : Akiyo, Bus, Football, Foreman et Soccer. Ce choix est justifié du fait que ces vidéos sont très utilisées pour ce genre de test, et elles couvrent différentes caractéristiques spatio-temporelles. Les quantités d'informations temporelles et spatiales sont décrites dans la section 2.5.2.2.1.

Pour toutes ces séquences, nous avons choisi le format CIF (352x288). Ensuite, les séquences vidéo sélectionnées sont codées en MPEG-2. Le taux d'image de ces séquences est de 30 images par seconde et le débit binaire moyen est de 300 kbit/s.

(a) Akiyo

(b) Bus

(c) Football



(d) Foreman

(e) Soccer

Figure 4.2. Séquences vidéos choisies

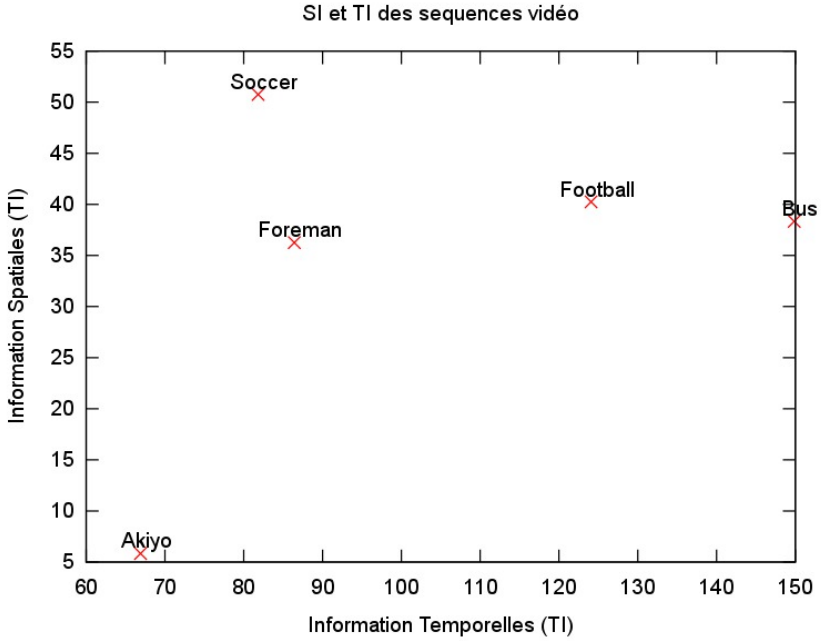


Figure 4.3. Quantités d'information temporelle et spatiale des séquences vidéo

Le serveur de diffusion est connecté à un ordinateur client par l'intermédiaire d'un réseau LAN. Nous avons utilisé le logiciel « *Darwin Streaming Server* » afin de diffuser les différentes séquences au client. Pour émuler les différents taux de perte de paquets, nous avons utilisé l'outil « *netem* », et plus particulièrement la commande *'tc'* pour avoir des pertes de paquet aléatoires.

Les séquences vidéo, reçues par le client, sont enregistrées pour pouvoir les comparer à la vidéo originale et de calculer leurs taux de ressemblance et de distorsion. Pour effectuer la comparaison entre la vidéo originale et la vidéo altérée, nous avons utilisé deux mesures objectives, qui sont le PSNR et le SSIM, et nous avons effectué un test subjectif pour avoir un score MOS de référence. Pour effectuer le test subjectif, nous avons demandé à un panel d'observateur (10 personnes) d'évaluer la qualité des vidéos altérées. Le test subjectif s'est déroulé selon les règles du standard ITU-R Rec. BT.500-11 [16], en appliquant la méthode « *Double Stimulus Impairment Scale* » décrite dans la section 2.5.1 de ce document.

Ensuite, nous avons utilisé les outils de *bSoft*, *Demokritos* et PSQA pour estimer la qualité perçue des vidéos altérées. Notez que, l'ensemble des paramètres utilisés par PSQA sont : le taux de perte de trame I, le taux de perte de trame P, le taux de perte de trame B, ainsi que la *Mean Burst Loss Size* (MBLS). Les valeurs mesurées sont ensuite transformées sur une échelle commune [0,10].

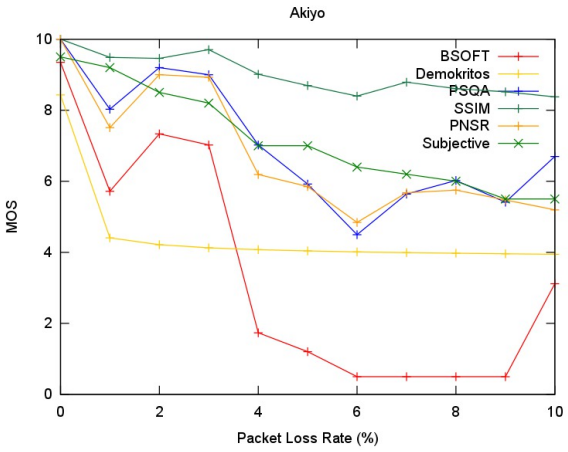
Cette version de PSQA a été proposée et étudié dans [74]. Le choix de ces paramètres (taux de perte des trames I, P et B) est justifié du fait que leur perte a un impact différent sur la qualité de la vidéo décodée. En effet, à cause des interdépendances des images du GOP, une erreur au niveau d'une image I se propage jusqu'à la prochaine image I. Les erreurs dans les images P se propagent jusqu'à la prochaine image I ou P. Les erreurs au niveau des images B ne se propagent pas.

4.2.2 Résultats

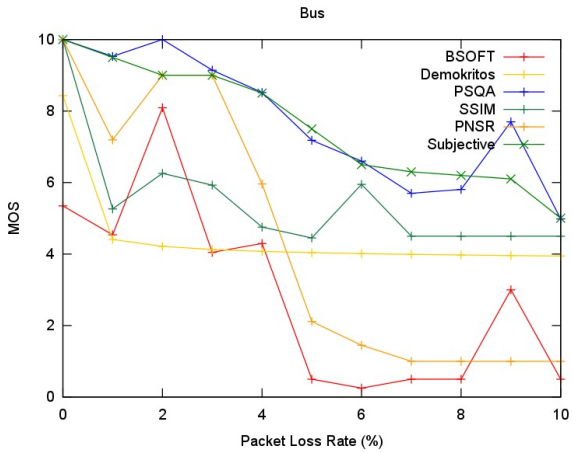
La Figure 4.4 (a, b, c, d et e) correspondent successivement à l'estimation du MOS des cinq séquences vidéo : Akiyo, Bus, Football, Foreman et Soccer. Le MOS est donnée par six différentes méthodes : PSQA, bSoft, Demokritos, PSNR, SSIM et un test subjectif (qui représentent la valeur de référence).

La première séquence vidéo (Akiyo) comprend un arrière-plan statique et une quantité de mouvement très faible : seul le visage de la présentatrice qui bouge. Par conséquent, la séquence est moins sensible aux pertes de paquet : la qualité vidéo reste acceptable même à 10% de perte. Presque toutes les méthodes d'évaluation ont donné des résultats proches du test subjectif.

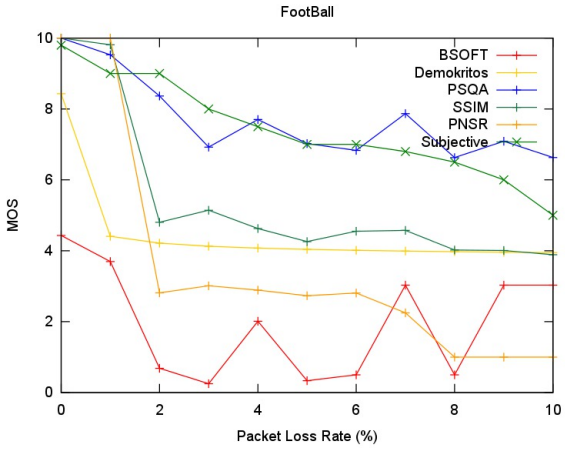
La séquence Foreman comporte le visage d'une personne avec des mimiques très riches. Il y a une quantité moyenne de mouvement, mais cette quantité de mouvement est un peu désordonnée et n'est pas homogène. En conséquence, la qualité de la vidéo diminue plus vite que la séquence Akiyo quand il y a beaucoup de pertes. Ainsi, la quasi-totalité des méthodes donne des mauvaises estimations de la qualité par rapport aux scores subjectifs, à l'exception de PSQA comme le montre la Figure 4.4(c).



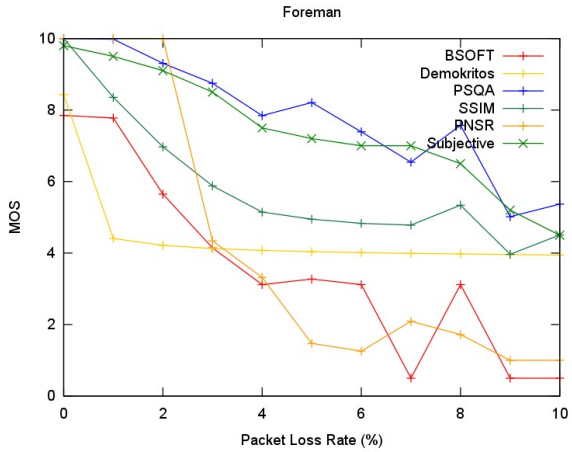
(a) Séquence Akiyo



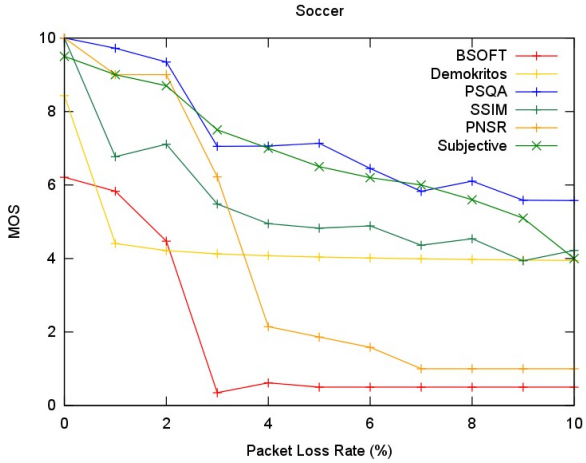
(b) Séquence Bus



(c) Séquence Football



(d) Séquence Foreman



(e) Séquence Soccer

Figure 4.4. Estimation du MOS

Par ailleurs, nous pouvons interpréter à partir de ces résultats que PSQA fournit une estimation (dans presque tous les cas de test) très proche de la valeur de référence, par rapport aux autres approches. Ceci est vrai pour toutes les séquences avec différentes quantités d'information spatiales et de quantité de mouvement (haute, moyenne et faible). PSQA est très robuste, en comparant aux autres approches, quand il y a une quantité considérable de mouvement dans la séquence vidéo.

Selon les courbes correspondantes aux séquences avec une quantité de mouvement moyenne et élevée, l'approche de bSoft et PSNR donnent une bonne estimation de la qualité quand les pertes ne dépassent pas les 3%. A partir de 3% de pertes, ces approches commencent à ne plus corréler avec les scores subjectifs, et considèrent que la séquence vidéo est de mauvaise qualité. Nous pouvons conclure que l'approche de bsoft et le PSNR sont plus adaptés pour la vidéo avec des mouvements faibles.

Contrairement à l'approche de bSoft, l'approche de Demokritos donne des mauvais scores quand il y a moins de 3% de pertes. A partir de 3% de pertes, l'approche de Demokritos donne des résultats similaires à ceux donnés par SSIM, mais toujours pas assez de corrélation avec les scores subjectifs.

Pour mieux étudier les performances des méthodes d'évaluation de la qualité vidéo, le groupe d'experts de la qualité vidéo (*Video Quality Expert Group* VQEG) a décrit différentes méthodes pour évaluer la performance des métriques de la qualité perçue [100]. Une de ces méthodes consiste à calculer le coefficient de corrélation de Pearson, qui donne des mesures quantitatives de la performance de la métrique proposée. Ce coefficient mesure la précision des estimations de la qualité vidéo par rapport aux résultats subjectifs.

Pour un ensemble de *N* paires de données (x_i, y_i), le coefficient de corrélation de Pearson est défini comme suit :

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad \textbf{(4.9)}$$

avec $\bar{x}$ et $\bar{y}$ sont respectivement des moyennes de *x* et de *y*.

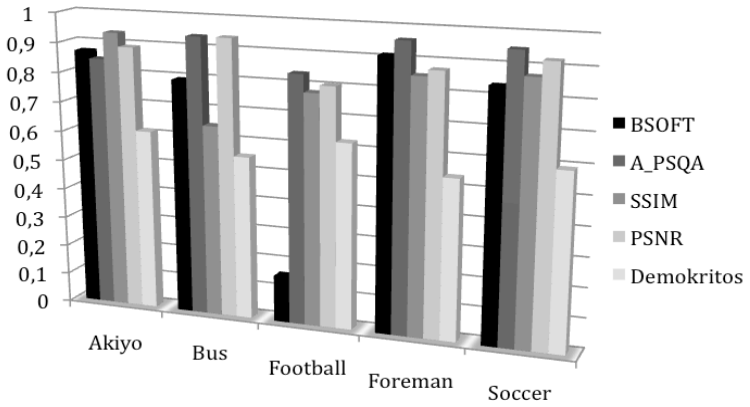


Figure 4.5. Coefficient de corrélation de Person

La Figure 4.5 montre que PSQA a une bonne corrélation avec les notes subjectives pour une variété de séquences vidéo (des séquences de faible activité telle que Akiyo et des séquences de hautes activités telles que bus et football).

Dans cette section, nous avons effectué un test subjectif pour étudier l'impact des paramètres réseaux sur la qualité perçue des vidéos. Nous avons remarqué que le paramètre réseau « perte de paquet » est le paramètre qui a le plus d'impact sur la qualité perçue. Les performances de plusieurs méthodes objectives avec et sans références ont été comparées. La méthode objective avec référence « SSIM » fournie des résultats proches du test subjectif, mais ne peut pas être utilisé en temps réel à cause du temps de calcul nécessaire pour l'estimation de la qualité. Nous avons aussi comparé différentes méthodes objectives sans référence. Les résultats montrent que la méthode PSQA, basée sur un réseau de neurones, fournie des bons résultats, très proches des scores des tests subjectifs. Nous rappelons que l'avantage des méthodes sans référence est qu'elles peuvent être utilisées dans un contexte temps réel.

Conclusion

Dans ce chapitre, nous avons présenté en détails la méthodologie PSQA. Cette méthodologie, basée sur les réseaux de neurones, fournie un bon outil d'estimation de la qualité d'un service.

Nous nous sommes intéressés à l'évaluation de la qualité des vidéos codées en MPEG-2. Nous avons évalué les performances de PSQA, en la comparant à d'autres méthodes d'évaluation objectives et subjective. Les résultats montrent que PSQA corrèle bien avec les évaluations subjectives pour différents types de vidéo et sous différentes conditions du réseau.

Chapitre 5

5 Évaluation des performances (en terme de qualité) de l'encodage et des encodeurs SVC

Le déploiement de SVC a un rôle important dans la diffusion des médias adaptatifs. En particulier, il permet l'adaptation au contexte de l'utilisateur et permet l'adaptation au niveau des réseaux émergents : *Content-Aware Networks* [81]. Les *Media Aware Network Elements* (MANEs) peuvent ajuster un flux SVC à la volée lors de la transmission sur réseau, afin de s'adapter aux nouvelles conditions du réseau (la congestion par exemple). Pour que cette technique fonctionne, le contenu doit être codé de façon appropriée, en tenant compte des capacités des terminaux (comme la résolution) et des caractéristiques du codec.

Ce chapitre élabore des recommandations de codage SVC pour le streaming média adaptatif des applications. Les performances « débit-distorsion » (*Rate-Distortion* RD) de ces recommandations sont validées pour différents codeurs. Plusieurs configurations d'encodage supplémentaires pour le streaming de médias adaptatif sont évaluées pour la haute définition (HD).

5.1 Recommandations d'encodage SVC

Parmi les solutions et les plates-formes les plus connues, nous trouvons : *Apple HTTP Live Streaming* (HLS), *Adobe HTTP Dynamic Streaming* (HDS), *Microsoft Smooth Streaming* (MSS), *YouTube* et *MTV* [85]... Plusieurs de ces technologies (à savoir HLS [82], HDS [83][84], MSS [85][86] et YouTube [87]) fournissent des recommandations pour l'encodage du contenu. Dans cette section, nous analysons brièvement ces recommandations et nous en déduisons des suggestions pour le streaming SVC.

Les résolutions énumérées dans ces recommandations, comme le montre le Tableau 5-1, vont de QCIF (176x144) à des débits d'environ 50 kbit/s jusqu'à 1920x1080 avec un débit maximum autour de 8 Mbps. En général, le nombre proposé de flux vari de un à quatre par résolution. Les résolutions communes, dans la plupart des plates-formes, sont 1280x720 et 1920x1080. Bien que la plupart des solutions de l'industrie concernent le déploiement du HTTP Streaming, les directives de codage que nous élaborons sont applicables aux médias SVC quel que soit la technique de transport réseau utilisée (UDP ou HTTP/TCP).

Résolution	Débit binaire [kbit/s]	Solution industrielle de streaming
1920x1080	6000, 5000	Microsoft Smooth Streaming
	8000, 6000, 5500, 5000, 4000	Adobe HTTP Dynamic Streaming
	7000-8000	Apple QuickTime
1280x960	4500	Apple HLS
1280x720	3450, 2272, 1672	Adobe Flash Media Server
	4000, 3500, 3000, 2500, 2000, 1500	Adobe HTTP Dynamic Streaming
	4500, 2500, 1800	Apple HLS
	5000-6000	Apple QuickTime
	3450, 3000, 2100, 1400	Microsoft Smooth Streaming
	2400 (live)	YouTube
	3500	MTV
960x540	2250	Microsoft Smooth Streaming
	1800	Apple HLS
	2200	MTV
720x486	1072, 672	Adobe Flash Media Server
854x480	1000 (live)	YouTube
848x480	1950	Microsoft Smooth Streaming
640x480	1200, 600	Apple HLS
	1000-2000	Apple QuickTime
848x440	1950	Microsoft Smooth Streaming
768x432	1740, 1140	Adobe Flash Media Server
	1700, 1500, 1200, 1000	Adobe HTTP Dynamic Streaming
	1700	MTV
736x416	1600	Microsoft Smooth Streaming
720x404	1500	Microsoft Smooth Streaming
640x360	1250	Microsoft Smooth Streaming

	1200, 600	Apple HLS
	600 (live)	YouTube
	1200	MTV
480x320	64	Apple HLS
554x304	950	Microsoft Smooth Streaming
400x300	400, 200, 110	Apple HLS
512x288	900	Microsoft Smooth Streaming
	650, 450, 300	Adobe Flash Media Server
	1700, 1500, 1200, 900, 600, 450, 300	Adobe HTTP Dynamic Streaming
	750	MTV
352x288	372, 268	Adobe Flash Media Server
448x252	450, 150	MTV
426x240	300 (live)	YouTube
416x234	400, 200, 110	Apple HLS
384x216	400	MTV
312x176	400	Microsoft Smooth Streaming
288x160	350	Microsoft Smooth Streaming
256x144	300, 250, 150	Adobe HTTP Dynamic Streaming
176x144	80, 32	Adobe Flash Media Server
	50-60	Apple QuickTime
112x64	50	Microsoft Smooth Streaming

Tableau 5-1. Suggestion de multiples débits binaire pour le streaming vidéo

Les solutions industrielles, citées au dessus, proposent des recommandations de codage : quel débit binaire il faut utiliser pour une certaine résolution. Sur la base de ces recommandations de codage, nous avons déduit une liste des résolutions typiques et les débits binaires correspondants pour le streaming SVC. Le Tableau 5-2 résume la liste des débits binaires à utiliser pour chaque résolution. Pour tenir compte des charges générales (en termes de donnée) du SVC, les débits binaires sont augmentés d'un certain pourcentage par rapport aux recommandations examinées. Dans la littérature, on suppose généralement un surcoût de codage de 10% de la couche d'amélioration par rapport à la simple couche AVC [88][89]. Pour un flux à 2 couches, nous proposons d'ajouter une surcharge de 10% pour les deux débits binaires. Cependant, pour les flux avec 4 couches, nous gardons le débit initial pour la couche de base afin de soutenir les faibles bandes passantes, et nous augmentons le débit de la première couche

d'amélioration de 10%, de la seconde couche de 20% et de la troisième couche de 30%. Par exemple, pour les débits binaires [4000 ; 5000 ; 6000 ; 8000], nous ajoutons la surcharge correspondante à chaque couche : [4000 ; 5000 + 10% * 5000 ; 6000 + 20% * 6000 ; 8000 + 30% * 8000]. En calculons les nouveaux débits binaires, qui prennent en compte les surcharges des couches d'amélioration, nous obtenons les débits binaires suivants : [4000 ; 5500 ; 7200 ; 10400].

Résolution	Débits binaires suggérés	
	4 débits binaires [kbit/s]	2 débits binaires [kbit/s]
1920x1080	10400, 7200, 5500, 4000	8800, 6050
1280x720	7800, 4800, 2750, 1500	5000, 2750
704x576	-	2200, 1350
960x540	-	2475, 1980
640x360	-	1760, 660
352x288	1950, 1080, 500, 270	1320, 330
176x144	-	110, 55

Tableau 5-2. Recommandations des taux binaires pour le streaming SVC

5.2 Banc d'essai

Quatre séquences vidéo en haute définition (1080p) ont été sélectionnées en fonction de leur information spatiale (SI) - c'est-à-dire, la quantité de détails spatiaux - et de l'information temporelle (TI) - c'est-à-dire, la quantité de mouvement (défini dans la section 2.5.2.2.1)- pour couvrir les différentes caractéristiques de contenu vidéo (cf. Figure 5.1) : PedestrianArea (SI bas, TI bas), Dinner (SI bas, TI haut), DucksTakeOff (SI haut, TI bas) et CrowdRun (SI haut, TI haut). La séquence Dinner a un taux de 30 images par seconde, les autres séquences ont un taux de 25 images par seconde. L'encodage s'est limité aux 250 premières images de chaque séquence vidéo. Nous avons testé l'encodeur AVC *x264* [90] et les principaux encodeurs SVC suivants : le codeur *Joint Scalable Video Model* (JSVM) [91], *MainConcept* [92], *VSS* [93] et *bSoft* [94]. Nos évaluations de RD sont basées sur *Peak Signal-to-Noise Ratio* (PSNR) et *NTIA Video Quality Metric* (VQM) [19].

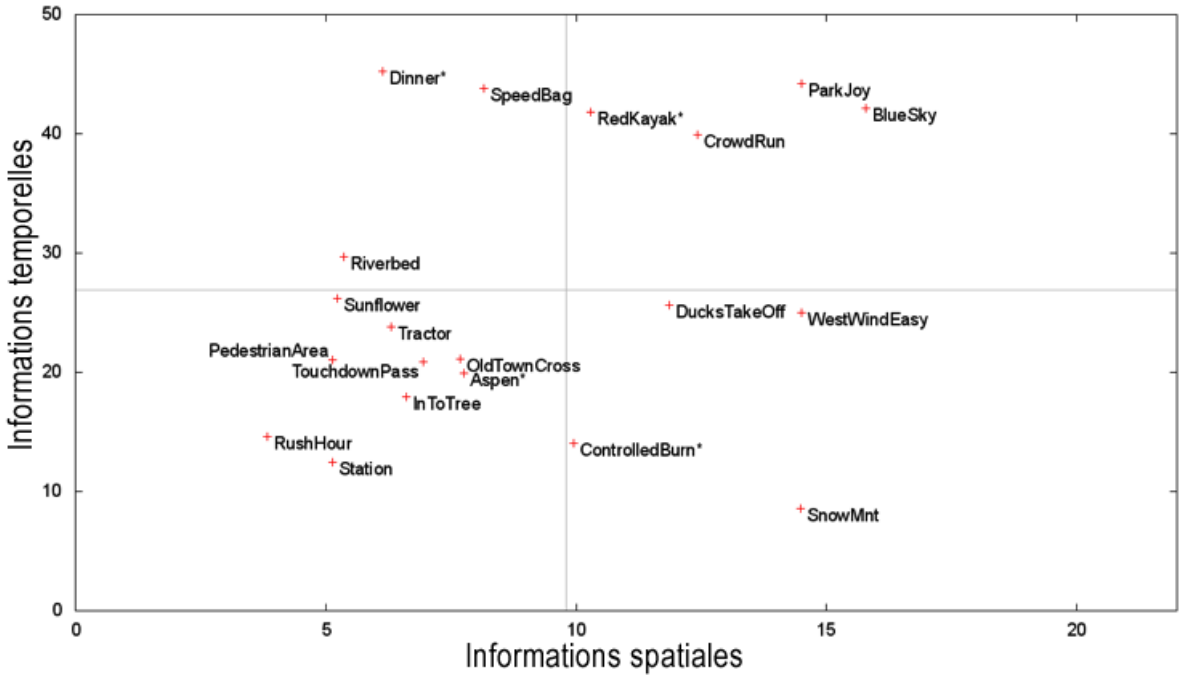
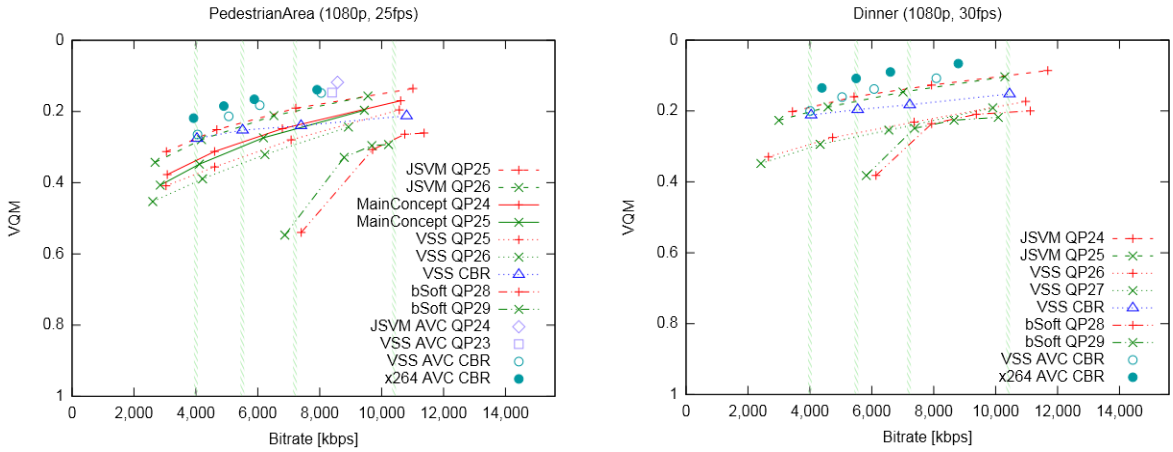


Figure 5.1. Informations spatiales et temporelles des vidéos connues

5.2.1 Performances des encodeurs

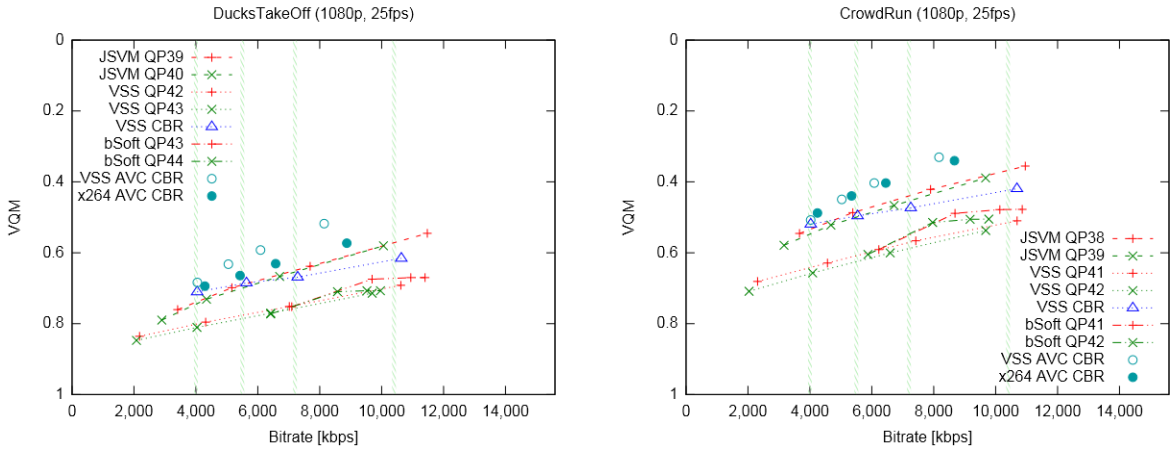
Nous commençons par comparer la performance RD de l'encodeur *x264* aux autres encodeurs SVC afin d'établir une ligne de base pour nos tests. Pour SVC nous utilisons une configuration à une seule couche (c'est-à-dire, une couche de base compatible AVC) et une configuration à 4 couches MGS (cf. section 2.3.4). Les flux vidéo avec une simple couche (AVC) sont codés en mode CBR avec les débits binaires suggérés pour le streaming dans le Tableau 5-2. Les flux SVC avec les 4 couches MGS sont codés en mode *Quantization Parameter* QP fixe pour tous les encodeurs SVC. La re-quantification entre les couches MGS a été fixé à un deltaQP égal à 2. Nous rappelons que le deltaQP représente la variation des paramètres de quantification QP entre deux couches MGS successives. Le deltaQP contrôle l'écart de qualité entre les couches consécutives. Plus le deltaQP est grand, plus il y a une différence de qualité remarquable entre les couches.

Les résultats VQM pour les séquences testées avec une résolution de 1080p sont présentés dans la Figure 5.2. Nous rappelons que les résultats VQM sont sur une échelle de valeur allant de 1 (forte distorsion) à 0 (pas de distorsion), indiquant la qualité d'une vidéo. Comme prévu, AVC donne une performance RD supérieure à SVC avec des couches MGS. Cependant, pour le débit binaire le plus faible, le flux SVC du codeur VSS dans le mode CBR (étiqueté VSS CBR) fournit la même qualité que l'AVC encodé avec le même codeur (étiqueté VSS AVC CBR).



(a) PedestrianArea

(b) Dinner



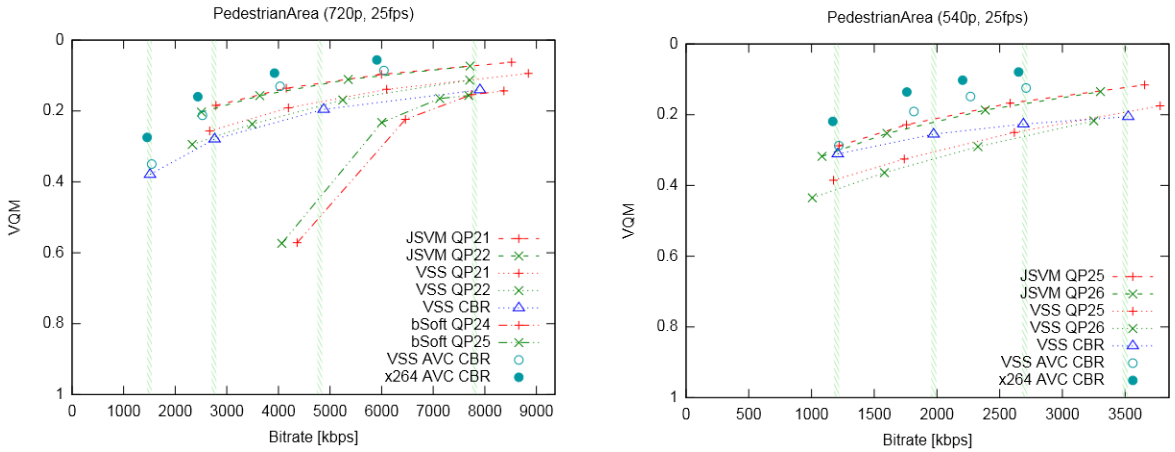
(c) DucksTakeOff

(d) CrowdRun

Figure 5.2. Résultats VQM du codage AVC et SVC avec différentes vidéos

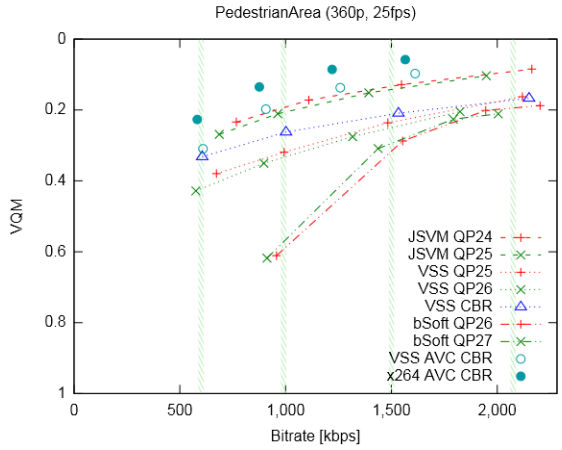
Le codeur JSVM avec les 4 couches MGS tend à atteindre la qualité des codeurs AVC dans plusieurs cas, mais nécessitant un peu plus de donnée (débit binaire supérieur). Parmi les codeurs SVC, le JSVM donne la meilleure performance RD, suivie en général par MainConcept, VSS et bSoft. Cependant, l'encodeur bSoft surpasse VSS pour des séquences plus complexes tels que CrowdRun.

Nous notons que le codeur VSS dans le mode CBR donne une qualité élevée à la couche de base, mais les couches d'amélioration SVC n'apportent (ou peu) pas d'amélioration à la qualité. Un débit plus élevé qui ne donne pas une meilleure qualité est essentiellement un gaspillage de bande passante, et donc n'est pas utile.



(a) 1280x720

(b) 960x540



(c) 640x360

Figure 5.3. Résultats VQM du codage AVC et SVC avec différentes résolutions

La Figure 5.3 montre les résultats VQM à des résolutions inférieures pour la séquence PedestrianArea. Comme avec 1080p, l'encodeur JSVM avec 4 couches MGS tend à atteindre la qualité du codage AVC dans la plupart des cas. En termes de besoins de stockage, SVC est plus efficace que AVC avec multiples représentations des 4 couches MGS, comme indiqué dans le Tableau 5-3.

Résolution	AVC	SVC	Réduction
1920x1080	23,000 kbit/s	10,400 kbit/s	54,8%
1280x720	14,000 kbit/s	7,800 kbit/s	44,3%
960x540	7,950 kbit/s	3,500 kbit/s	55,8%
640x360	4,350 kbit/s	2,080 kbit/s	52,2%

Tableau 5-3. Stockage des flux SVC par résolution

5.2.2 Résultats et discussion : impact de *deltaQP*

Nous évaluons, dans cette section, les performances d'encodage des codeurs SVC pour une seule résolution spatiale (1920x1080) avec quatre couches MGS et des *deltaQP* (dQP) variables entre les couches. Les scores PSNR et VQM de toutes les séquences de test sont présentés dans la Figure 5.4, Figure 5.5, Figure 5.6 et la Figure 5.7. Dans ces figures, le type des lignes indique l'encodeur utilisé et la couleur de la ligne indique la valeur de dQP. Notez que, contrairement aux autres codeurs qui utilisent la re-quantification pour les couches MGS, le codeur bSoft distribue automatiquement les coefficients de transformation à travers les couches [95], ce qui élimine la nécessité d'encoder avec différents dQP dans ce test.

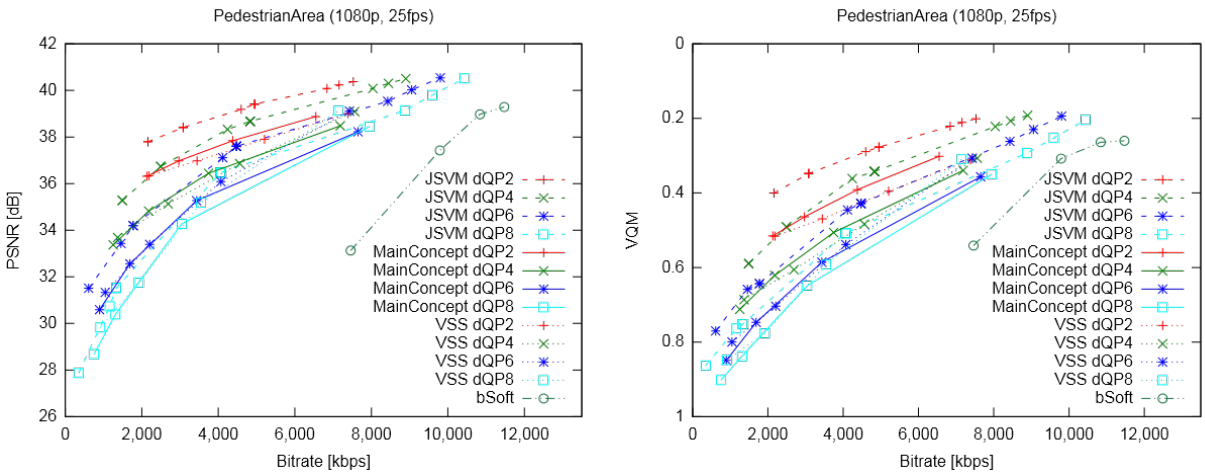
Nous pouvons observer qu'un dQP égale à 2 est suffisant pour offrir une bonne gamme de débits binaires avec 4 couches, tout en ayant la meilleure performance RD. De toute évidence, les valeurs dQP élevées (par exemple 6 ou 8) provoquent une forte quantification de la couche de base, ce qui induit une mauvaise qualité (scores VQM proches de 1), comme le montre la Figure 5.4. L'encodeur JSVM présente les meilleures performances RD dans toutes les séquences de test.

Comme les encodeurs industriels SVC sont optimisés pour avoir une forte vitesse d'encodage, ils sacrifient généralement un peu de performance RD, par exemple, en utilisant des algorithmes rapides de recherche de bloc d'estimation de mouvement. D'autre part, l'encodeur JSVM utilise des calculs complexes tout au long de la chaîne d'outils d'encodage vidéo pour assurer une performance RD élevée.

L'encodeur bSoft distribue automatiquement les coefficients de transformation pour créer des couches MGS. La couche de base donne des résultats (PSNR) assez faibles et le *bitstream* a des débits binaires nettement plus élevés que les autres encodeurs pour le même PSNR. L'encodeur bSoft fournit des résultats égaux à ceux des autres encodeurs en termes de débit-distorsion.

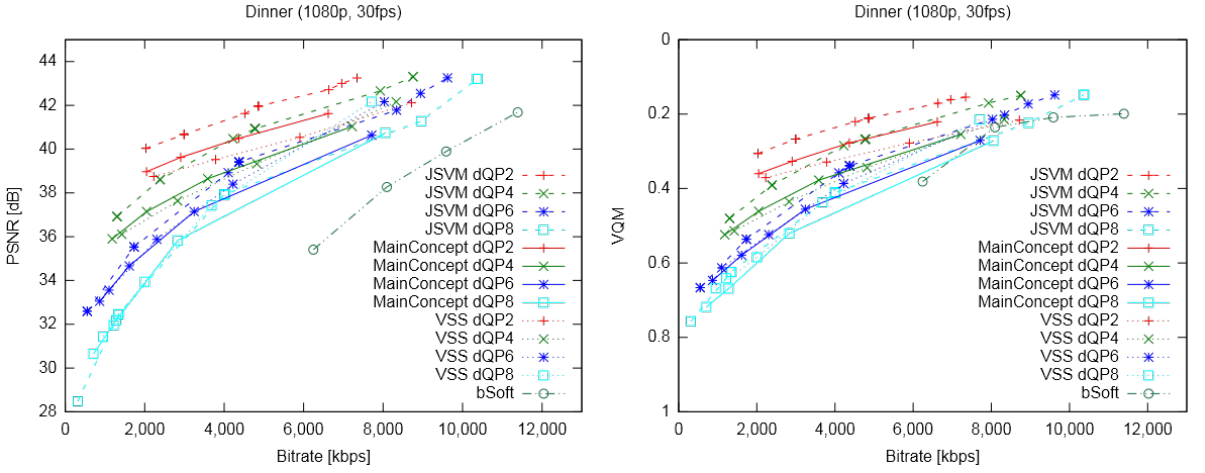
Pour mieux analyser les résultats, nous avons comparé les scores PSNR vs VQM pour les quatre séquences, comme indiqué dans la Figure 5.8. La figure montre clairement que, pour un même PSNR, les couches inférieures, encodé avec bSoft, ont des scores VQM meilleurs que les autres encodeurs. Pour les autres encodeurs, les résultats de la Figure 5.8 montrent une forte corrélation entre PSNR et VQM, à part quelques exceptions au niveau des couches inférieures. Notez toutefois que cette corrélation est dépendante du contenu.

Afin d'étudier l'impact de la variation de dQP sur les performances de codage, la Figure 5.9 montre les résultats de l'encodeur JSVM, dans le cas où le facteur de quantification QP de la couche la plus haute est égale à 28. Nous observons que dQP = 2 est généralement suffisant pour offrir une bonne qualité avec des débits binaires adéquats. Plus nous avons un dQP faible, plus les performances RD sont meilleurs, bien que l'effet diminue pour des scènes plus complexes (avec une plus grande SI et/ou TI).



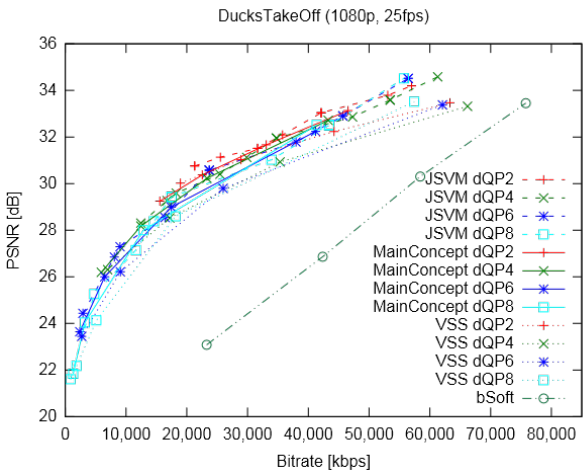
(a) PSNR (b) VQM

Figure 5.4. Variation du dQP entre les couches MGS pour différents encodeurs
Séquence *PedestrianArea*

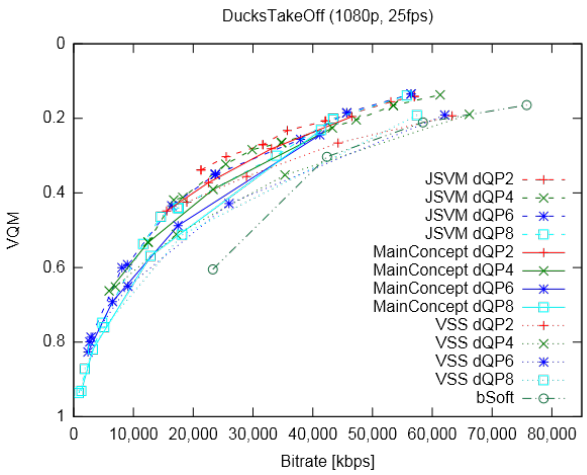


(a) PSNR (b) VQM

Figure 5.5. Variation du dQP entre les couches MGS pour différents encodeurs
Séquence *Dinner*

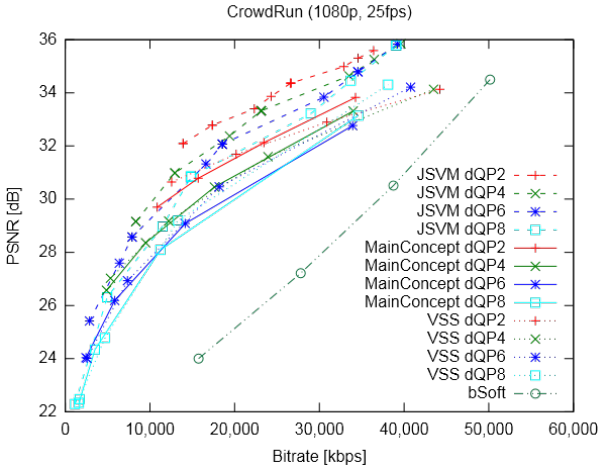


(a) PSNR

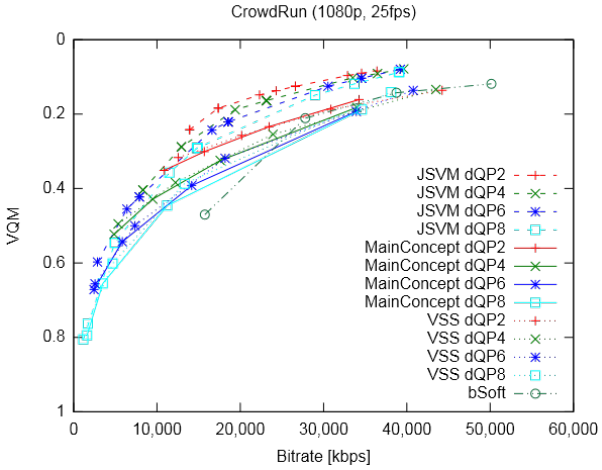


(b) VQM

Figure 5.6. Variation du dQP entre les couches MGS pour différents encodeurs
Séquence *DucksTakeOff*

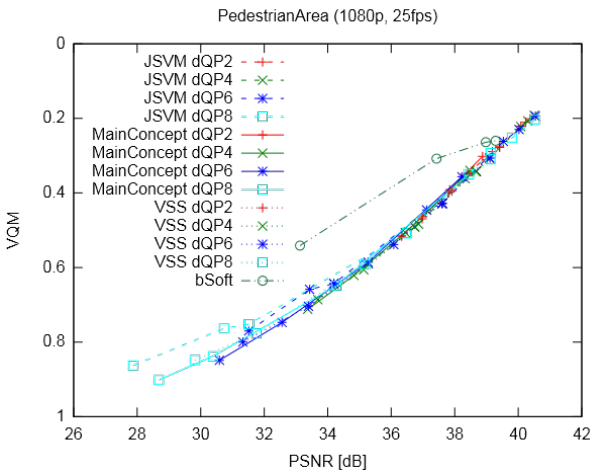


(a) PSNR

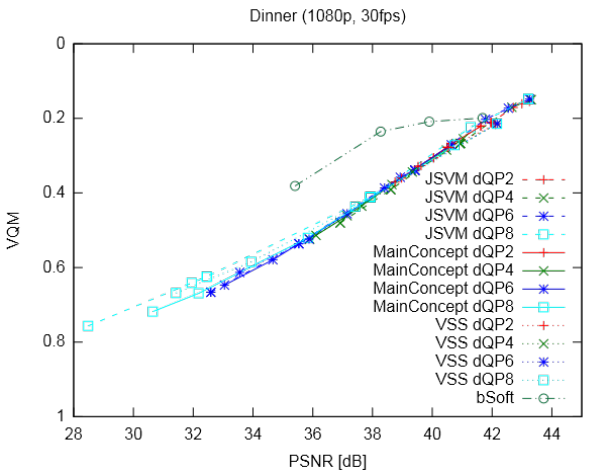


(b) VQM

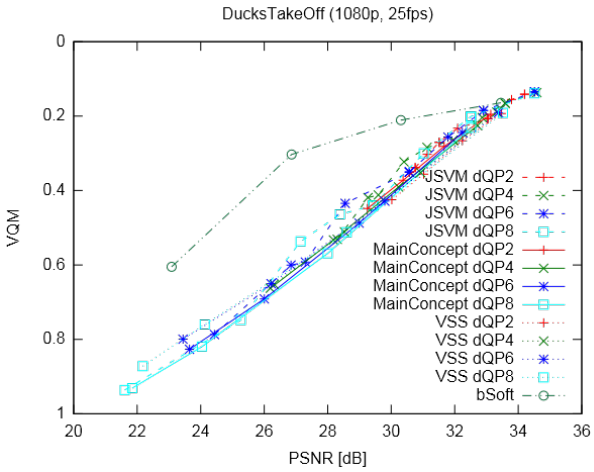
Figure 5.7. Variation du dQP entre les couches MGS pour différents encodeurs
Séquence *CrowdRun*



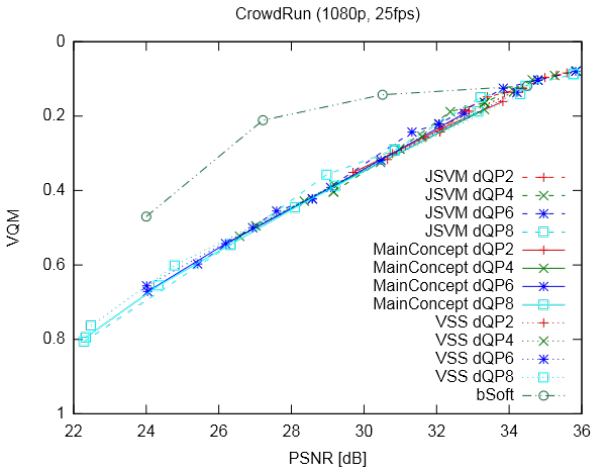
(a) PedestrianArea



(b) Dinner

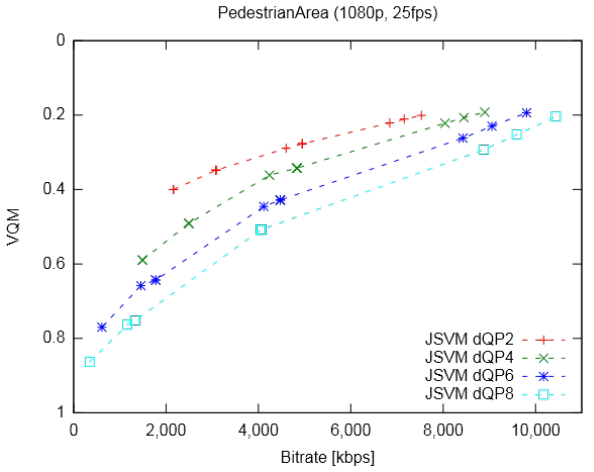


(c) DucksTakeOff

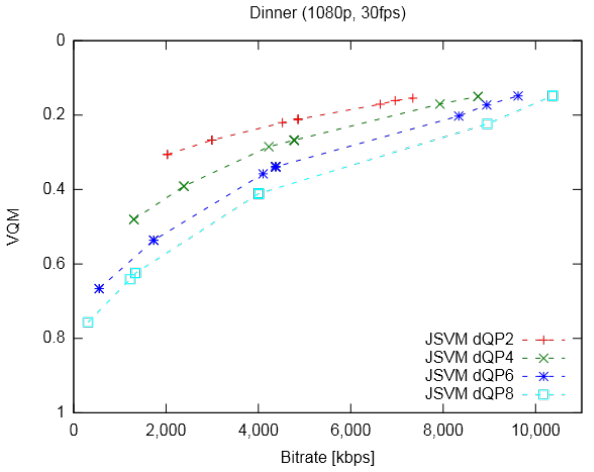


(d) CrowdRun

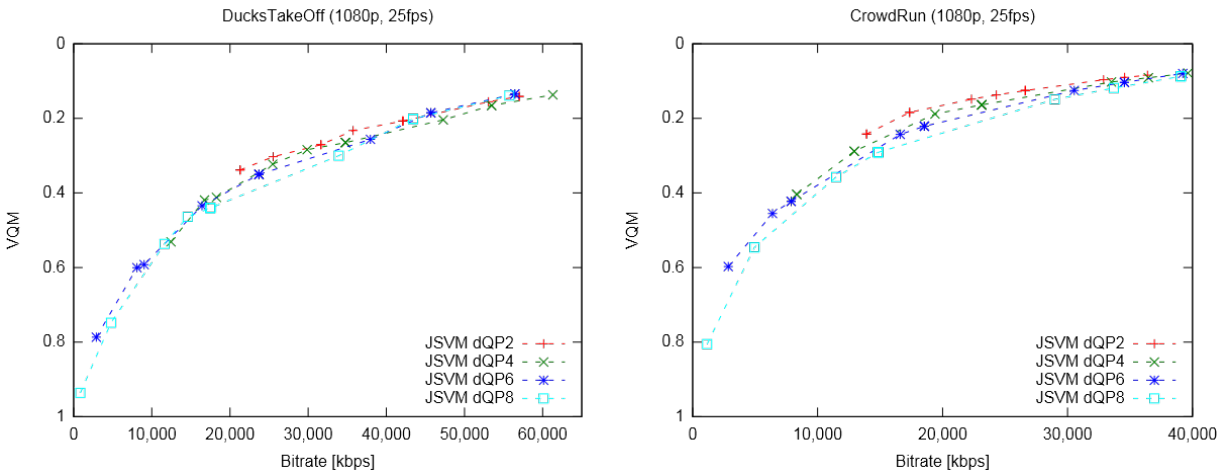
Figure 5.8. Corrélation entre PSNR et VQM pour différents dQP des couches MGS pour différents encodeurs



(a) PedestrianArea



(b) Dinner



(c) DucksTakeOff (d) CrowdRun

Figure 5.9. Résultats VQM en variant dQP entre les couches MGS pour l'encodeur JSVM

Les durées d'encodage par trame sont présentées dans la Figure 5.12. En raison d'une vitesse d'encodage très faible de JSVM, les résultats sont représentés sur une échelle logarithmique. Les encodeurs MainConcept et VSS sont les plus rapides, et ils sont presque deux cents fois plus rapide que JSVM. L'encodeur bSoft est vingt fois plus rapide que JSVM. Pour tous les encodeurs, les vitesses d'encodage sont un peu plus lentes pour des faibles valeurs de dQP. En outre, la séquence Dinner bénéficie d'une plus courte durée d'encodage dans tous les encodeurs, probablement parce que c'est une scène synthétique, suivi par PedestrianArea.

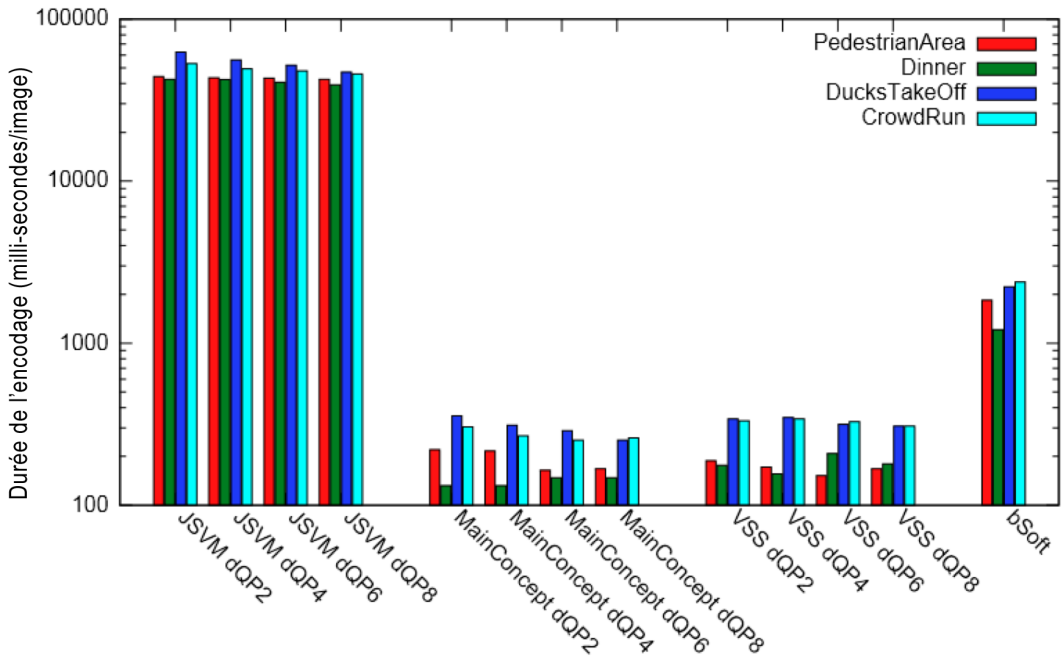


Figure 5.10. Durée de l'encodage avec différents dQP entre les couches MGS pour différents encodeurs

5.2.3 Résultats et discussion : CGS vs. MGS

Comme mentionné dans la section 2.3.4, il existe deux approches pour la scalabilité de la qualité dans SVC : *Coarse Gain Scalability* (CGS) et *Medium Grain Scalability* (MGS). CGS crée essentiellement plusieurs couches spatiales, avec la même résolution, tant dis que MGS partitionne les coefficients de transformation d'une image codée, de façon à obtenir des qualités différentes d'une vidéo.

La performance RD de l'encodeur bSoft pour les couches MGS et CGS est illustrée dans la Figure 5.11. Nous avons comparé les points d'extraction de flux SVC pour la séquence CrowdRun avec 4 couches MGS contre 4 couches CGS avec deltaQP égale à 2. Notez encore une fois que les QPs correspondent à la couche SVC la plus haute.

Les résultats montrent que MGS a une meilleure performance RD au niveau de la couche la plus haute, mais pour les couches inférieures, les performances RD se dégradent rapidement. D'autre part, CGS maintient une performance RD constante, bien que les débits soient plus élevés. Pour les couches SVC inférieures, la performance RD du CGS est généralement meilleure que pour MGS pour un même débit. Les résultats PSNR de couches de base en mode CGS varient entre 22,64 dB à 4,771 kbit/s et 28,12 dB à 12 543 kbit/s. Pour le mode MGS, les résultats PSNR des couches de base sont relativement stables (entre 24,00 et 24,22 dB), indépendamment de la valeur du QP. En d'autres termes, la qualité de base reste la même, quelle que soit la qualité que nous voulons atteindre pour la couche la plus élevée. L'encodeur met beaucoup d'informations au niveau de la couche de base pour la prédiction des couches supérieures. Ces informations augmentent le débit binaire mais n'ont pas d'impact sur la qualité de la couche de base.

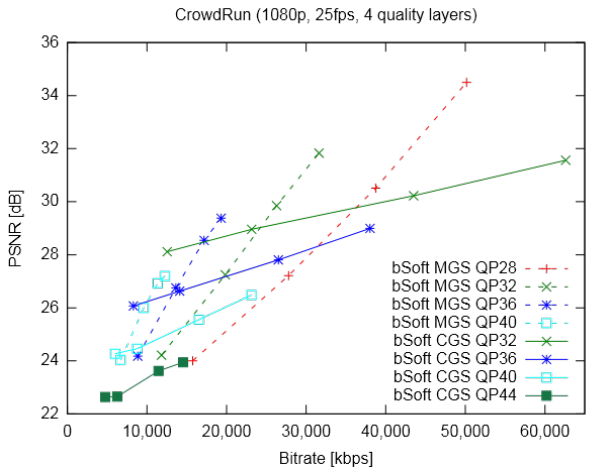


Figure 5.11. Résultat PSNR avec l'encodeur bSoft – MGS vs CGS

5.2.4 Résultats et discussion : Nombre des couches MGS

Le test suivant étudie l'impact du nombre de couches SVC en mode MGS sur la performance RD. Intuitivement, un grand nombre de couches engendre des pénalités au niveau du débit binaire. Nous avons testé les codeurs JSVM, MainConcept et bSoft avec la configuration suivante : Le paramètre QP de la plus haute couche a été fixé à 28. Pour le codeur MainConcept, dQP a été fixé à 2.

La Figure 5.12 montre les résultats PSNR de 1 à 4 couches MGS pour les codeurs JSVM, MainConcept et bSoft. Les codeurs JSVM et MainConcept présentent une diminution assez constante des

performances RD quand le nombre de couches augmente. Les résultats de l'encodeur bSoft montrent un écart important entre le débit d'une seule couche de codage (étiqueté bSoft 1MGS sur la Figure 5.12) et les *bitstreams* avec de multiple couches MGS. Les *bitstreams* avec 2 et 3 couches MGS (étiqueté respectivement par bSoft 2MGS et bSoft 3MGS) ont presque la même performance RD. La couche de base de bSoft 2MGS a même un débit et un PSNR inférieur que ceux de la couche de base de bSoft 3MGS.

Les résultats PSNR des plus hautes couches restent relativement stables à travers le nombre de couches MGS pour les différents codeurs. Par contre, les codeurs allouent moins de qualité pour les couches de base pour chaque couche supplémentaire MGS.

En moyenne, le codeur JSVM nécessite un débit supplémentaire d'environ 17% pour ajouter une couche MGS. Le codeur MainConcept nécessite un débit d'environ 20% de plus, et le codeur bSoft nécessite un débit binaire supplémentaire d'environ 6%.

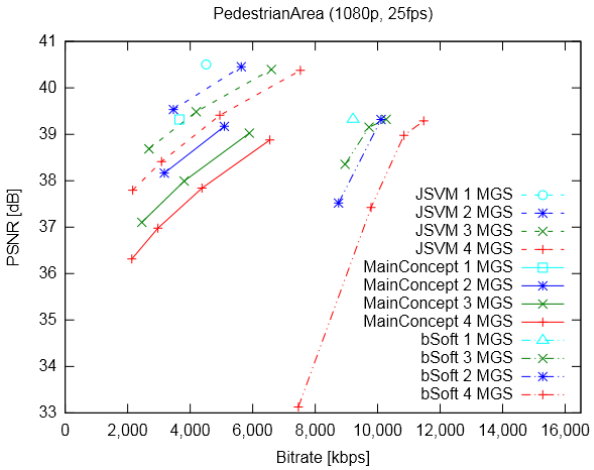


Figure 5.12. Variation du nombre de couche MGS

Conclusion

Nous avons étudié dans ce chapitre les performances des encodeurs SVC les plus connues pour le streaming média MPEG-AVC/SVC et nous avons proposé des débits binaires pour l'encodage SVC. Nous avons validé ces choix et nous avons proposé certaines configurations de codage des contenus vidéo en haute définition. Par ailleurs, nous avons testé et mis en évidence les caractéristiques des différents codeurs.

Nos évaluations montrent que le mode CBR ainsi que les modes de contrôle de débit à QP fixes, fournissent une très bonne qualité pour les débits binaires suggérés pour toutes les résolutions. Nos résultats indiquent également que c'est plus adapté, pour les streaming médias, d'encoder un flux SVC par résolution plutôt que d'un seul flux comprenant toutes les résolutions.

Chapitre 6

6 Evaluation de la qualité des flux vidéo SVC

Il existe plusieurs approches pour mesurer la qualité de la vidéo : évaluations subjectives et des évaluations objectives (cf. section 2.5). Pour résumé, les évaluations subjectives sont constituées d'un ensemble d'humain qui note la qualité des séquences vidéo de courte durée en fonction de leur propre point de vue personnel. Les évaluations objectives sont principalement des algorithmes et des formules qui permettent de mesurer la qualité dans un système automatique, de manière quantitative et reproductible. Plusieurs méthodes objectives ont été proposées dans la littérature. La majorité de ces méthodes prend la vidéo d'origine comme référence, ou du moins comme une référence partielle, pour estimer la qualité d'expérience finale. Toutefois, les méthodes objectives avec référence pourraient être coûteuses et inutilisable en temps réel, puisque la vidéo originale est requise. Pour résoudre cette contrainte, des modèles basés sur des fonctions empiriques, ont été proposés (tel que présenté dans [96][97]).

Détaillé dans la section 2.3.4, le codec H.264/SVC a été finalisé en Mai 2003 et se propose d'augmenter l'efficacité du codage et de la compression tout en facilitant l'adaptation du flux vidéo aux caractéristiques du réseau de transport. Cette adaptation est facilitée par l'introduction d'une nouvelle couche dans l'architecture du standard, appelée *Network Abstraction Layer* (NAL) qui génère des unités appelées *NAL Units* (NALUs). Concrètement, cette variante du MPEG-4 Part 10 (H.264/AVC) permet en fait d'intégrer les données nécessaires à plusieurs niveaux de décodage dans le même flux. Avec ce système, le streaming est plus simple : au lieu de proposer un choix de fichiers à l'utilisateur en fonction de ses capacités, c'est le fichier qui s'adapte à l'utilisateur.

Dans le chapitre précédent, nous avons étudié les performances des encodeurs SVC et les configurations SVC qui fournissent les meilleures performances RD. Nous avons utilisé PSNR et VQM, deux outils objectifs et avec référence, pour évaluer la qualité des flux SVC. Dans ce chapitre, nous allons vérifier l'impact que peut avoir certains paramètres réseau et encodeur, sur la qualité perçue. Suite à cette évaluation de la qualité, nous allons proposer deux méthodes objectives et sans référence pour estimer la qualité perçue en temps réel.

6.1 Méthode proposée

La méthode proposée pour estimer la qualité d'expérience de flux vidéo SVC, est basée sur les réseaux de neurones et est similaire à l'évaluation de la qualité pseudo-subjective (PSQA) proposé dans [101]. L'idée principale est d'avoir plusieurs échantillons déformés évalués subjectivement, puis d'utiliser les résultats de cette évaluation pour entrainer à un réseau de neurones la relation entre les paramètres qui provoquent la distorsion et la qualité perçue. La procédure consiste à choisir un ensemble de paramètres

qui ont un impact significatif sur la qualité perçue. Ces paramètres peuvent appartenir à la métrique de qualité de service du réseau (QoS) et/ou aux paramètres d'encodage vidéo. Une base de données d'échantillons déformés est alors générée, en faisant varier les paramètres d'encodage vidéo et les conditions de la transmission vidéo du réseau.

Normalement, tous les échantillons déformés doivent être évalués subjectivement. Comme mentionné dans la section 2.5.1, les évaluations subjectives nécessitent beaucoup de temps et d'effort. En plus, nous sommes en disposition des vidéos originales (non altérées). Par conséquent, nous avons décidé d'utiliser une méthode d'évaluation objective appelée NTIA VQM, qui a une forte corrélation avec les scores d'évaluation subjective [21].

Pour chaque paire de séquence vidéo (original et dégradé), l'algorithme de NTIA fournit une métrique d'estimation de la qualité, avec des valeurs comprises entre 0 et 1 (0 lorsqu'il n'y a pas de différences perçues et 1 pour la dégradation maximale) qui peuvent être directement liés à la DMOS (*Differential Mean Opinion Score*). Les valeurs DMOS renvoyées par le modèle NTIA peuvent être liée à la MOS en utilisant l'équation (6.1). L'interprétation des valeurs MOS est présentée dans le Tableau 6-1.

$$MOS = 5 - 4*DMOS \qquad \textbf{(6.1)}$$

MOS	Qualité	VQM
5	Excellent	< 0,2
4	Bien	> 0,2 & < 0,4
3	Moyen	> 0,4 & < 0,6
2	Mauvais	> 0,6 & < 0,8
1	Médiocre	> 0,8

Tableau 6-1. Correspondance entre les scores VQM et MOS

Après la procédure d'évaluation de la qualité par VQM, une partie de la base de données des échantillons déformés et leurs scores VQM correspondants sont utilisés pour l'apprentissage du réseau de neurones. La partie restante de la base de données est utilisée pour la validation et le test du réseau de neurones.

Comme décrit ci-dessus, la première étape la plus importante dans notre méthode, est le choix des paramètres affectant la qualité. Nous devons choisir les paramètres qui ont le plus d'impact sur la qualité perçue, et qui peuvent être mesuré en temps réel. Il y a beaucoup de facteurs qui peuvent influer sur la qualité des flux vidéo SVC. Les facteurs dépendent des conditions du réseau (paramètres QoS tels que la perte de paquets, bande passante ...), de l'application (IPTV, VoD, streaming mobile ...) et des paramètres d'encodage vidéo (résolution, paramètres de quantification, le nombre de couches, etc.). En outre, le choix des paramètres devrait être applicable à tous les types de scalabilité (qualité, temporelle, spatiale, ou une combinaison de celles-ci).

6.1.1 Les paramètres d'encodage vidéo affectant la qualité

Dans un flux vidéo H.264/SVC, les paramètres de codage vidéo ont un impact évident sur la qualité vidéo perçue. Les paramètres les plus importants, à priori, pour la vidéo sont la résolution vidéo, la cadence des images et le paramètre de quantification QP. Des études ont été menées (comme dans [15], [16] et [17]) afin d'étudier l'impact de ces paramètres sur la qualité vidéo perçue dans des contextes différents. Le résultat de ces études confirme, comme prévu, que ces paramètres ont un impact important sur la qualité subjective de la vidéo numérique. Dans nos recherches, nous allons nous limiter à une résolution fixe qui est de 1280x720, et nous allons varier la cadence des images (7,5, 15 et 30 images par seconde).

La relation entre les paramètres de quantification QP et de la qualité estimée par VQM est montrée dans la Figure 6.1. La baisse des valeurs du paramètre de quantification QP conduit à une meilleure qualité vidéo. Par exemple, une valeur de QP inférieur à 30 fournit une excellente qualité vidéo, mais une valeur de QP supérieur à 44 conduit à une mauvaise qualité vidéo.

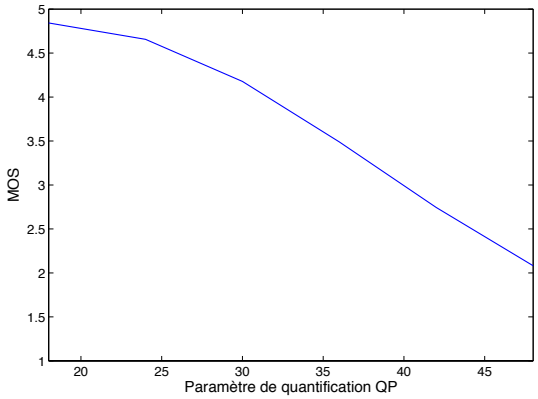


Figure 6.1. Impact du paramètre de quantification sur la qualité vidéo

6.1.2 Les paramètres réseau affectants la qualité : taux de perte des NALU

Nous rappelons que NALU (*Network Abstraction Layer Unit*) est l'unité de transport de paquets vidéo. L'en-tête des NALUs indique le type et « l'importance » du NALU. L'extension d'en-tête a des champs particuliers notés D, Q et T, qui sont respectivement utilisés pour identifier les couches spatiales, les couches de qualité et les couches temporelles. La charge utile des NALUs contient des informations de commande ou des données vidéo codées.

Un NALU ne peut transporter que l'information d'une couche particulière, par conséquent, une perte de NALU ne touche que la couche appropriée. Nous insistons sur le fait que la perte d'un NALU appartenant à la couche de base a plus d'impact sur la qualité de la vidéo, que de la perte de NALU appartenant à d'autres couches d'amélioration, comme a été prouvé dans [101]. Ceci est dû au fait que les images des couches d'amélioration sont déduites à partir de celles de la couche de base.

6.2 Expérimentations et résultats

6.2.1 Test avec VQM

Pour simplifier le réseau de neurones, et d'accroître son efficacité, nous avons choisi les paramètres suivants : le paramètre de quantification minimale de toutes les couches et le taux de perte NALU de chaque couche (*Base Layer, Enhancement Layer 1, Enhancement Layer 2*).

Afin d'entrainer le réseau de neurones pour estimer la qualité de la vidéo, nous avons considéré trois séquences vidéo différentes : Aspin, RedKayak et RushFieldCuts. Les vidéos ont été encodées en utilisant l'encodeur JSVM [98]. Nous avons utilisé le logiciel openSVC [99] du côté du décodeur. Les vidéos sont encodées sur 3-couches avec une scalabilité SNR et le mode CGS, avec une résolution 1280x720 à 30 images par seconde. Le paramètre de quantification QP varie de 18 à 48 dans les 3 couches, avec $QP_{\text{couche de base}} > QP_{\text{couche de renforcement 1}} > QP_{\text{couche de renforcement 2}}$. Les pertes NALU ont été générées à partir de 0% jusqu'à 20% sur chaque couche. La chaîne de Markov à deux états (modèle simple Gilbert-Elliot) est utilisée pour la génération des processus de pertes.

En prenant en compte la combinaison des paramètres cités ci-dessus et leurs valeurs, nous générons environ 200 séquences vidéo altérées. La qualité de toutes les vidéos présentant une distorsion a été évaluée à l'aide de VQM. L'ensemble des données obtenues ont été divisés en trois groupes : 70% des données relatives à l'entrainement du réseau de neurones, 15% pour la validation et 15% pour les tests. Nous avons utilisé une structure de réseau de neurones à 3-couches (*feed-forward*) en utilisant *PSQA*. L'architecture du réseau de neurones est décrite dans la Figure 6.2. La première couche du réseau de neurones dispose de 4 neurones correspondant aux 4 entrées (QP_{min}, le taux de perte de la couche de base, le taux de perte de EL1, taux de perte de EL2). La couche cachée contient 10 neurones. Le nombre de neurone dans la couche cachée est choisi, après plusieurs simulations, de façon à obtenir les meilleurs résultats. La couche finale est composée d'un seul neurone pour la sortie. La sortie du réseau de neurones correspond à l'estimation de la qualité.

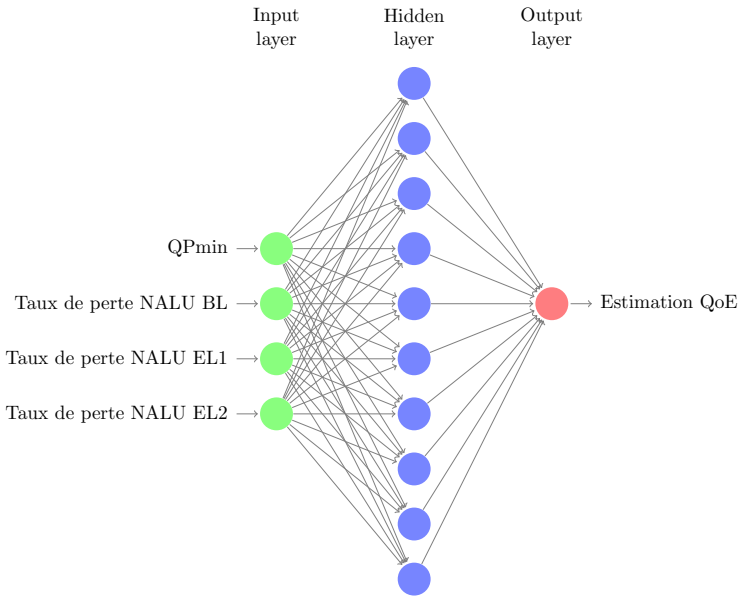


Figure 6.2. Architecture du réseau de neurones

Figure 6.3 montre comment la qualité de la vidéo est sensible aux pertes NALU de la couche de base et au paramètre de quantification. En fait, la qualité de la vidéo estimée diminue quand QP augmente ou lorsque le taux de perte des NALU de la couche de base augmente.

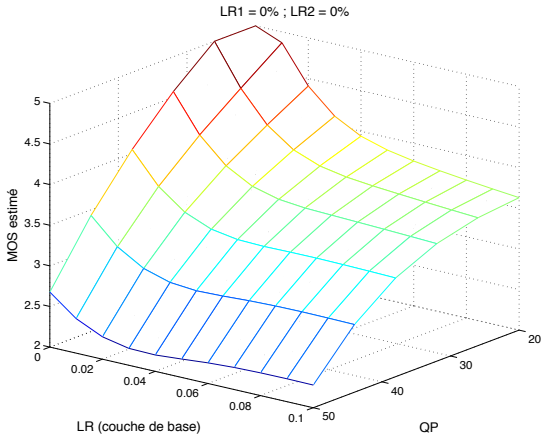


Figure 6.3. MOS estimé en fonction du QP et du taux de perte de la couche de base

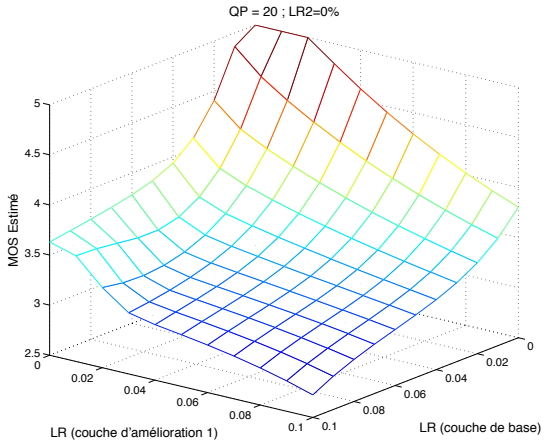


Figure 6.4. MOS estimé en fonction des taux de perte de la couche de base et de la couche d'amélioration 1

Comme représenté sur la Figure 6.1 et confirmé dans la Figure 6.3, la qualité de la vidéo est acceptable lorsque le paramètre QP est inférieur à 38. Dans le cas d'un QP supérieur à 40, la qualité vidéo est médiocre, surtout quand il y a des pertes de NALU. Comme prévu, la qualité de la vidéo est plus sensible aux pertes NALU de la couche de base que ceux d'une couche d'amélioration, comme le montre la Figure 6.4. En effet, la couche de base est plus sensible aux pertes NALU, par rapport aux autres couches, parce que toutes les couches d'amélioration utilisent les images de la couche de base comme référence et toute erreur dans la couche de base se propagerait aux autres couches. En outre, dans le cas de perte dans les couches d'amélioration uniquement, le décodeur a une meilleure chance de corriger les erreurs, par rapport au cas où il y a des pertes de NALU dans la couche de base. Nous pouvons voir sur la Figure 6.5 que la qualité vidéo est relativement un peu plus sensible à une augmentation des pertes NALU dans EL1 comparé aux pertes NALU dans EL2.

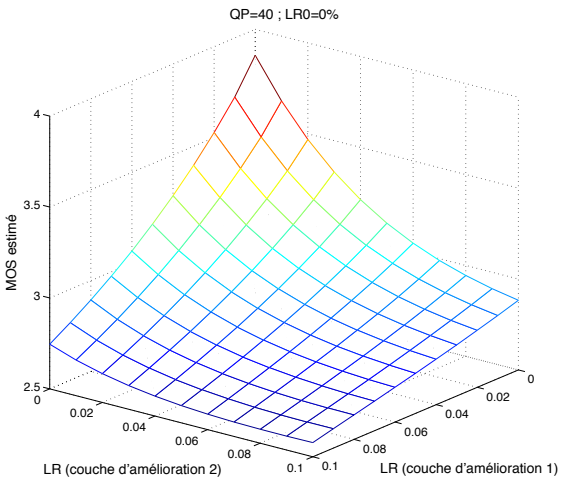


Figure 6.5. MOS estimé en fonction des taux de perte des couches d'amélioration 1 et 2

Afin d'évaluer les performances de notre méthode, et de valider la corrélation avec les scores VQM, nous calculons l'erreur quadratique moyenne (MSE) et le coefficient de Pearson (*R*). MSE quantifie la différence entre les valeurs données par VQM et les valeurs estimées par le réseau de neurones proposé. MSE doit être inférieure à 10^{-2} pour une bonne estimation. Le coefficient de Pearson mesure la corrélation entre les qualités vidéo estimées par VQM et par le réseau de neurones proposé. Plus le coefficient est proche de 1, plus forte est la corrélation. De façon générale, un facteur de corrélation supérieur à 0,8 est considéré comme assez élevé.

Le MSE (*Mean-Square-Error*) obtenu est égale à 0,0529, ce qui signifie qu'il y a une différence minime entre les scores VQM et les scores estimés. Le coefficient de corrélation *R* est égal à 0,9209, ce qui reflète une très bonne corrélation.

La Figure 6.6 montre le diagramme de dispersion des scores VQM contre les scores estimés avec le réseau de neurones proposé. Plus les points coïncident avec la ligne tracée en diagonale, plus forte est la précision de l'outil proposé. Comme représenté sur la Figure 6.6, le nuage de points est proche de la ligne tracée en diagonale, ce qui valide la bonne corrélation entre VQM et le réseau de neurones.

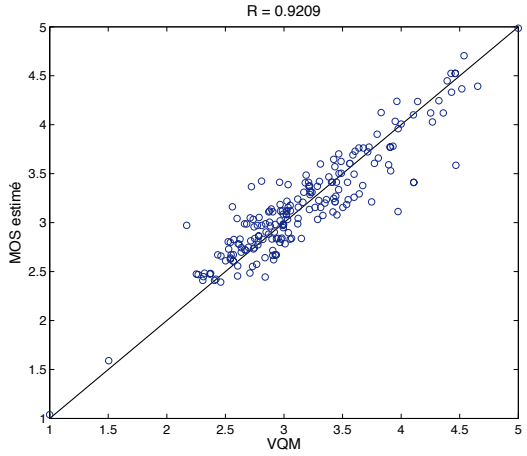


Figure 6.6. Corrélation entre VQM et MOS estimé

6.2.2 Test avec une évaluation subjective

Le test décrit dans la partie précédente a été réalisé avec des séquences vidéo à 30 images par seconde. Ceci est principalement dû aux limitations de VQM. En effet VQM compare les deux séquences vidéos image par image et donc ne prend pas en compte le taux d'image affiché aux observateurs, alors que c'est un paramètre important vis-à-vis de la perception des utilisateurs. Pour cette raison, nous avons décidé d'effectuer un test subjectif selon les normes pour voir quel impact pourrait avoir les paramètres choisis, combinés avec le changement du taux d'image, sur la qualité perçue.

6.2.2.1 Déroulement du test subjectif

Pour la réalisation du test subjectif, quatre vidéos différentes ont été utilisé : Aspin, RedKayak, RushFieldCuts et ControlledBurn. Les vidéos ont été encodé avec différentes configurations : nombre de couche de 1 à 3, taux d'image de 7,5 fps, 15 fps ou 30 fps, variation de QP pour les différentes couches, variation des taux de perte des NALU.

En variant ces paramètres, nous obtenons un grand nombre de vidéo à évaluer par chacun des observateurs. Le problème c'est que l'évaluation d'un grand nombre de vidéo, même de courte durée (10 secondes), est très fatigante pour les observateurs. C'est pour cette raison que nous avons décidé de faire une sélection d'échantillons les plus pertinents afin de réduire le nombre de vidéo à évaluer. Après la sélection, nous avons obtenu 130 séquences vidéo altérées, avec des caractéristiques différentes.

Le test subjectif a été réalisé selon les normes citées par l'ITU [17]. Les séquences vidéos sont affichés par pair (vidéo non altérée, puis vidéo altérée) et d'une façon aléatoire. Dix observateurs notent la qualité perçue de chaque séquence vidéo. La méthode de notation utilisée pour le test subjectif est « *double-stimulus impairment scale* » (Recommandation ITU-R BT.500 [16]). Les scores fournis par les observateurs sont sur une échelle de 1 à 5. Le score 1 est pour une dégradation perceptible et très gênante, et le score 5 est pour une vidéo de très bonne qualité, sans dégradation perceptible.

Pour étendre notre étude, nous avons aussi utilisé VQM pour mesurer la qualité d'une façon objective. La Figure 6.7 montre la différence entre les scores obtenus par VQM et les MOS (scores obtenus avec le test subjectif). Nous remarquons que l'outil VQM ne corrèle pas assez bien avec les scores subjectifs quand le taux d'image est égal à 7,5 ou à 30 images par seconde. Par contre, dans le cas des vidéos à 15 images par seconde, VQM fournit des scores relativement proches du MOS. En effet, les observateurs sont plus sensibles aux taux d'image affiché. Un taux de 7,5 images par seconde est un peu gênant pour les observateurs, ce qui résulte des scores relativement faibles par rapport à ceux obtenus par VQM. Par contre, les utilisateurs semblent être plus satisfaits (par rapport aux scores de VQM) quand le taux est de 30 images par seconde.

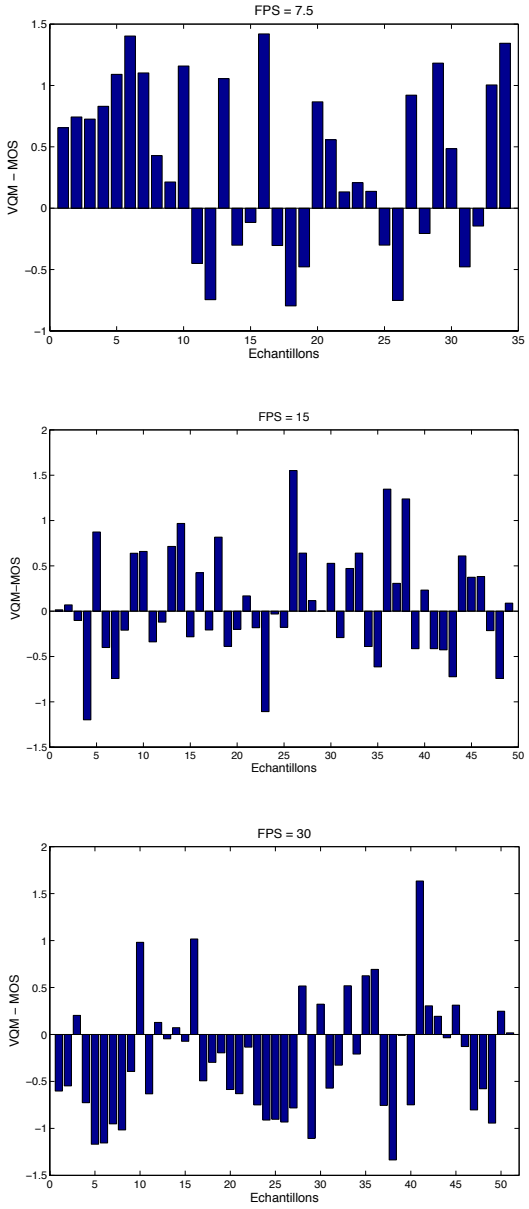


Figure 6.7. Différence entre VQM et MOS (FPS=7,5 ; 15 ; 30)

6.2.2.2 Entrainement des réseaux de neurones

Comme il a été précisé précédemment, les paramètres (variables) qui ont un impact sur la qualité sont : le nombre de couche du flux SVC, le taux d'image par seconde, le QP de la couche la plus haute, le taux de perte des NALUs de chaque couche et le MLBS de la couche de base. Pour simplifier le réseau de neurones (minimiser le nombre d'entrée), et gagner en précision, nous avons décidé de générer neuf réseaux de neurones différents : un réseau de neurones pour chaque combinaison (nombre de couche, taux d'image). Nous rappelons que le nombre de couche varie de 1 à 3, et que le taux d'image peut avoir une valeur égale à 7,5, 15 ou 30 images par seconde.

Pour entrainer un réseau de neurones à estimer le MOS dans le cas d'un flux SVC à une couche, nous choisissons comme paramètres d'entrées : (1) facteur de quantification QP, (2) le taux de perte des

NALUs et (3) la taille moyenne des pertes consécutives des NALUs. La Figure 6.8 montre l'impact du MLBS et des pertes des NALUs de la couche de base. Le MOS estimé décroit quand la taille moyenne des pertes consécutives des NALUs augmente, et comme prévu, les distortions sont rapidement perceptibles dès qu'il y a des pertes de NALUs. La qualité est jugée mauvaise à partir de 6% de perte de NALUs.

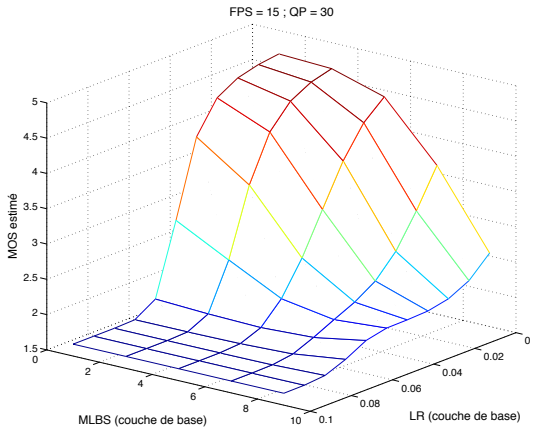


Figure 6.8. Estimation du MOS dans le cas d'un flux SVC à 1 couche (FPS=15 ; QP=30)

Dans le cas d'un flux SVC à deux couches à 15 fps, les entrées du réseau de neurones sont : (1) le facteur de quantification, (2) le taux de perte des NALUs de la couche de base, (3) la taille moyenne des pertes consécutives de la couche de base et (4) le taux de perte des NALUs de la couche d'amélioration 1.

Comme le montre la Figure 6.9, les pertes des NALUs ont un impact important sur la qualité. Le réseau de neurones se comporte d'une façon logique, vu que le MOS estimé se dégrade rapidement sous l'effet des pertes des NALUs.

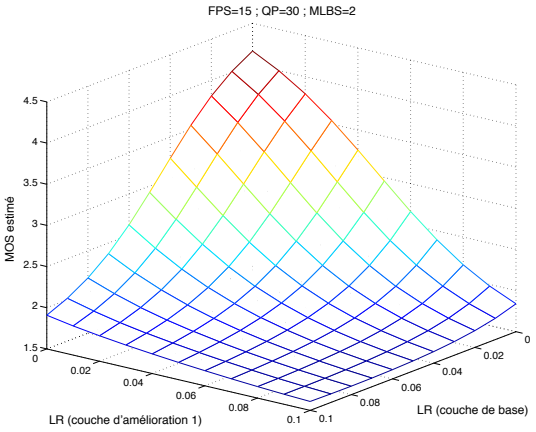


Figure 6.9. Estimation du MOS dans le cas d'un flux SVC à 2 couches (FPS=15 ; QP=30 ; MLBS=2)

Les réseaux de neurones, dans le cas d'un flux SVC à trois couches, ont tous les mêmes entrées, et donc le même nombre de neurone d'entrée, comme indiqué dans la Figure 6.10.

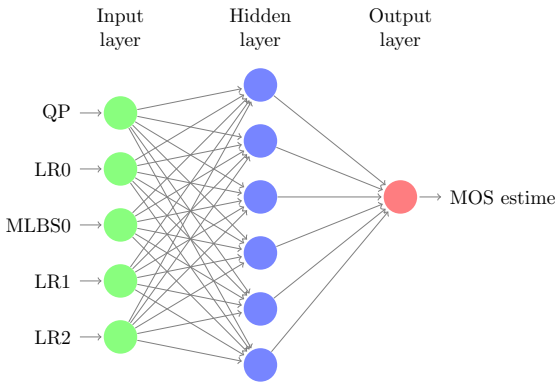
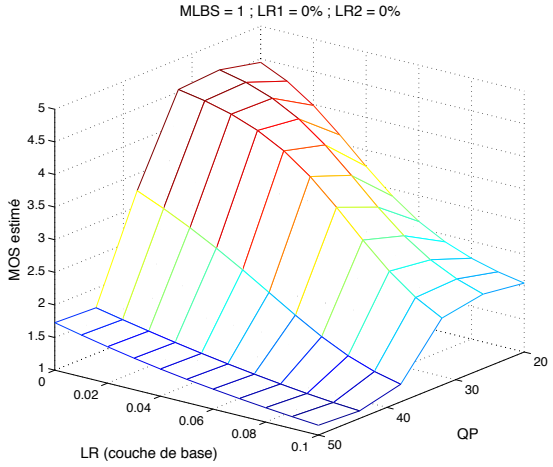
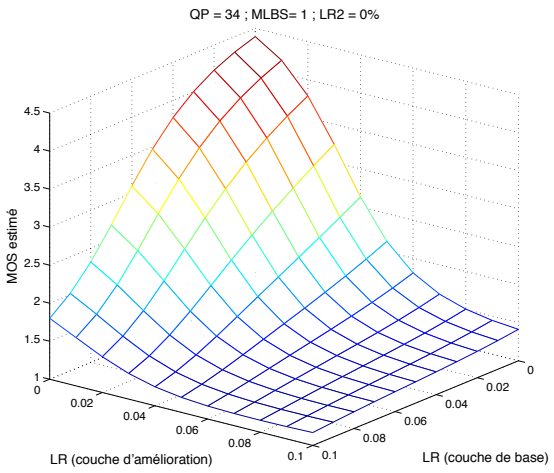


Figure 6.10. Architecture du réseau de neurones pour 3 couches

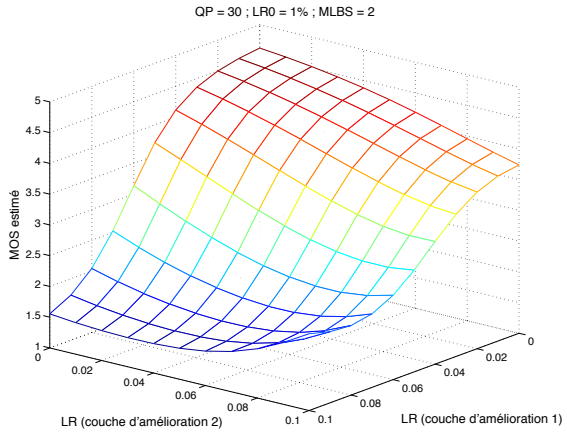
La Figure 6.11 montre les résultats obtenus dans le cas du réseau de neurones entrainé pour les flux SVC à trois couches et avec un taux d'image de 15 images par seconde. Le réseau de neurones se comporte d'une façon cohérente. En effet, nous remarquons qu'il y a un impact différent des pertes de NALU des différentes couches et du facteur QP. La qualité se dégrade dès que le facteur QP dépasse la valeur 35 quand il n'y a pas de perte de NALU. La qualité se dégrade aussi rapidement quand il y a des pertes de NALUs au niveau de la couche de base. Dans la majorité des cas, les imperfections commencent à être perceptibles par les observateurs à partir de 4% de perte de NALUs. Quand le taux de perte de NALUs dépasse les 4%, les imperfections deviennent très gênantes pour les observateurs. Le réseau de neurones reflète bien ces cas de figure et fournit des scores, dans la majorité des cas, inférieurs à 3 quand le taux de perte des NALUs dépasse les 4%.



(a) MOS estimé en fonction de QP et du taux de perte de la couche de base



(b) MOS estimé en fonction des taux de perte de la couche de base et de la couche d'amélioration 1



(c) MOS estimé en fonction des taux de perte des couches d'amélioration 1 et 2

Figure 6.11. MOS estimé en fonction des paramètres d'entrer

Pour confirmer les performances du réseau de neurones, la Figure 6.12 montre la corrélation entre les résultats du test subjectif (MOS), les résultats fournis par les réseaux de neurones (avec différents taux d'image) et l'estimation de la qualité par VQM. Nous remarquons que les scores fournis par le réseau de neurones (représentés par l'axe des coordonnées) sont très proches de la diagonale, ce qui signifie qu'ils sont très proches des scores du test subjectif (représentés sur l'axe des abscisses). Les scores estimés par VQM sont un peu éloignés de la diagonale. Le facteur de corrélation *R* du réseau de neurones est égal à 0,99, ce qui reflète une très bonne corrélation du réseau de neurones avec les tests subjectifs, alors que celui de VQM est égale à 0,86.

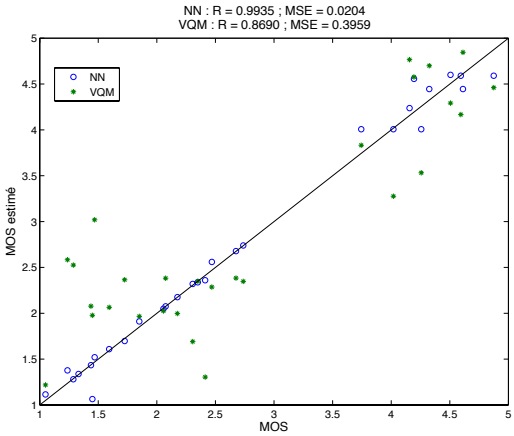


Figure 6.12. Corrélation entre les scores du test subjectif (MOS) et le MOS estimé

Conclusion

Dans ce chapitre, nous avons étudié l'impact, que peut avoir certains paramètres réseaux et paramètres liés à l'encodeur, sur la qualité perçue des vidéos SVC et nous avons proposé deux outils objectifs sans référence pour l'estimation de la qualité. Pour cela, nous avons utilisé deux méthodes différentes pour évaluer la qualité : une méthode objective, en utilisant l'outil VQM et une méthode subjective. Les premiers résultats montrent que la qualité perçue par les utilisateurs est sensible aux taux d'image, aux facteurs de quantification et aux taux de perte des NALU.

Afin de pouvoir évaluer la qualité d'une façon objective, et sans vidéo de référence, nous avons entrainé plusieurs réseaux de neurones pour les différentes caractéristiques des vidéos SVC. Nous avons proposé un réseau de neurones basé sur les évaluations objectives par VQM. Le réseau de neurones obtenu fournit une bonne estimation des scores VQM. En comparant les résultats fournis par VQM et les résultats des tests subjectifs, nous avons remarqué que VQM ne corrèle pas bien avec les MOS dans le cas des taux d'image 7,5 et 30 images par seconde, et donc il vaut mieux entrainer le réseau de neurones en se basant sur les résultats du test subjectif.

Dans la dernière partie, nous nous sommes basé sur les résultats du test subjectif pour entrainer un réseau de neurones. L'estimation du MOS avec le réseau de neurones corrèle très bien avec le test subjectif et fournit des meilleurs résultats que VQM.

Il est important de rappeler que la méthode proposée, basée sur un réseau de neurones, prend parti de la précision des tests subjectifs, et peut être utilisée dans un contexte temps réel, sans avoir besoin de la séquence vidéo de référence.

Chapitre 7

7 Evaluation de la qualité de la Voix sur IP

Dans ce chapitre, nous allons proposer différentes méthodes pour l'évaluation de la qualité d'une communication VoIP. Nous proposons une première méthode basée sur la régression polynomiale et une deuxième méthode basée sur la méthodologie PSQA. Dans un premier lieu, nous allons étudier l'impact de plusieurs paramètres (réseau et applicatif) sur la qualité perçue des séquences vocales. Par la suite, nous proposons une première méthode basée sur l'interpolation polynomiale et une deuxième méthode basée sur les réseaux de neurones.

7.1 Méthode proposée

Les méthodes proposées sont des méthodes non-intrusives et paramétriques qui permettent l'estimation de la QoE. L'idée principale est d'avoir plusieurs échantillons déformés évalués subjectivement, puis d'utiliser les résultats de cette évaluation pour entrainer un réseau de neurones aléatoire, ou trouver une interpolation polynomiale.

Comme décrit dans le chapitre 4, la première étape consiste à identifier les paramètres qui ont une incidence sur la qualité d'expérience dans un contexte donné. Pour le cas de la VoIP, parmi les paramètres qui peuvent avoir le plus d'impact sur la qualité perçue par les utilisateurs, nous citons les paramètres réseaux tels que les pertes de paquets et la taille moyenne des pertes consécutives. Les pertes de paquet introduisent des coupures dans une communication VoIP, et ceci pourrait être gênant aux interlocuteurs. La taille moyenne des pertes consécutives est utile pour avoir une idée sur la longueur des durées des interruptions. En fait, les longues interruptions sont facilement perceptibles pour les utilisateurs, même si le taux global de perte de paquets reste relativement faible. Ces longues interruptions sont très gênantes pour les interlocuteurs.

La gigue et le délai dans le cas d'une communication VoIP peuvent aussi avoir un impact pertinent sur la qualité. En fait, si nous n'utilisons pas une taille adéquate des buffers et les bons algorithmes de routage et de priorité, la gigue et le délai vont causer l'élimination de certains paquets qui arrivent en retard à l'utilisateur. C'est pour cette raison que les paquets éliminés à cause de la gigue et du délai seront considérés comme perdus.

Une fois que les paramètres ont été choisis (pourcentage de perte de paquets et la taille moyenne des pertes consécutives), nous définissons un intervalle et un ensemble de valeurs représentatives pour chacun d'eux, selon certaines conditions de fonctionnement du système de test. Le nombre de valeurs à choisir pour chaque paramètre dépend de la taille de l'intervalle choisi et de la précision souhaitée. Dans ce contexte, nous appelons « une configuration » l'ensemble des valeurs choisies pour chaque paramètre.

Une base de données de séquences vocales altérés est alors générée par la simulation d'une communication VoIP, en utilisant l'ensemble des configurations possibles.

La qualité des séquences de voix altérées est ensuite évaluée. La méthode d'évaluation la plus précise de la qualité de la voix est la méthode subjective. Les méthodes d'évaluation subjectives sont précises et donnent une estimation réelle de la qualité perçue par les interlocuteurs. L'inconvénient des méthodes subjectives est qu'elles ne peuvent pas être utilisées d'une façon automatique, et qu'elles sont très couteuses en terme de temps, surtout quand il y a un grand nombre de séquence à évaluer.

Par conséquent, nous nous sommes retourné vers les méthodes d'évaluation objectives, parce que : (i) nous sommes en possession des séquences originales, (ii) l'évaluation de la qualité se fait de manière automatique et rapide, (iii) les méthodes intrusives, spécialement la méthode PESQ décrite dans la section 2.5.2.3.3, ont une bonne corrélation avec l'évaluation subjective. L'ensemble des séquences vocales est ainsi évalué objectivement, en utilisant l'outil PESQ. Il résulte de cette évaluation, une base de données contenant les différentes configurations des paramètres *P* et les scores objectifs correspondants (que nous notons MOS-like).

Des polynômes et des réseaux de neurones sont ensuite utilisés pour reproduire cette correspondance entre les paramètres appliqués à chaque séquence et le MOS-like correspondant. Certaines des configurations sont utilisées pour l'entrainement du réseau de neurones et les configurations restantes sont utilisées pour la validation du réseau de neurones obtenu. Une fois que le réseau de neurones est entrainé et validé, nous obtenons un outil qui permet d'évaluer la qualité à partir des paramètres choisis, sans avoir besoin de la séquence vocale originale.

7.1.1 Environnement de test

Afin de former une base de données, de séquences vocales altérées, complète, nous avons utilisé six différents échantillons de parole standards (d'une durée de 10 secondes). Les séquences originales ont été échantillonnées à une fréquence de 8 kHz, codées sur 16 bits. La moitié d'entre elles étaient des voix masculines et l'autre moitié des voix féminines.

Pour simuler les communications VoIP, nous avons utilisé l'outil PJSIP [75]. PJSIP est une bibliothèque gratuite et Open Source de communication multimédia. C'est une API de communication multimédia de haut niveau qui convient à presque tous les types de systèmes allant des ordinateurs de bureau, les systèmes embarqués, les téléphones portables. PJSIP supporte plusieurs codecs voix, y compris les plus récents.

Nous avons décidé de travailler avec les codecs les plus connus et les plus utilisés : iLBC, Speex et Silk (codec de Skype). Ces codecs sont relativement récents, et fournissent une bonne qualité vocale lors des communications VoIP. Nous avons limité l'utilisation de ces codecs en mode bande étroite (fréquence d'échantillonnage de 8 kHz) en raison de la limitation de l'outil PESQ pour l'évaluation de la qualité.

Comme décrit dans la section précédente, nous avons pris le taux de perte de paquets (LR) et la distribution de ces pertes (MLBS) comme paramètres influents sur la qualité de la voix.

Pour simuler les pertes de paquets, plusieurs auteurs proposent différents modèles mathématiques. Les modèles mathématiques varient du simple (par exemple, en supposant que les pertes sont indépendantes et uniformément réparties dans le temps), au plus complexes (par exemple des chaînes de

Markov d'ordre *n*). Le modèle de Markov à deux états (modèle de Gilbert simple) est largement utilisé dans la littérature [76][77] pour simuler les pertes de paquets, car c'est un modèle précis et simple pour obtenir des processus de pertes comme ceux que nous trouvons sur Internet. Comme indiqué dans la section 3.3.1.3, qui fournit plus de détail sur le modèle de Gilbert simple, l'expression de p et q en fonction des paramètres LR et MLBS, est la suivante :

$$p = \frac{1}{MLBS} \cdot \frac{LR}{1 - LR} \; ; \; q = \frac{1}{MLBS} \qquad (7.1)$$

Nous avons considéré les valeurs de taux de perte de 1% à 30% avec un pas de 1%. La taille moyenne des pertes consécutives varie de 1 à 7. Chaque paquet transporte *20 ms* de Voix. Certaines combinaisons de LR et MLBS ne sont pas réalisables à cause de la durée limitée des séquences vocales (10 secondes). Ainsi, seules les combinaisons valides ont été prises en compte. Pour chaque combinaison de LR et MLBS, nous avons généré dix processus de perte (modèle de Gilbert) différents.

Les codecs, que nous avons choisis, utilisent le masquage de perte (*Packet Loss Concealment*), dans le cas de perte de paquets, et la détection d'activité de la voix (*Voice Activity Detection*) pour ne pas envoyer une grande quantité de donnée en cas de silence des interlocuteurs. Pour que la méthode proposée soit efficace, nous avons décidé d'étudier les qualités obtenues dans le cas où le masquage de perte est activé et dans le cas où le masquage de perte n'est pas activé. Concernant le *Voice Activity Detection*, nous tenons à souligner que les pertes sont imperceptibles quand ils se produisent au cours d'une période de silence, et donc les processus de perte de paquets ne s'appliquent pas dans le cas où il y a un silence.

Les séquences vocales altérées sont ensuite évaluées, d'une façon automatique, en utilisant l'outil PESQ. Nous obtenons une estimation de la qualité pour chaque séquence vocale, et pour chaque configuration. Pour les descripteurs statistiques de l'ensemble des valeurs PESQ, d'une configuration donnée (codec, état PLC, LR, MLBS), nous avons choisi d'utiliser la médiane, et non la moyenne, ni la variance. La médiane est en fait une bonne approximation des scores PESQ comme indiqué dans [78]. Nous avons donc appliqué la médiane pour les 10 différents processus de perte et pour les 6 séquences de voix pour chaque quadruplet (Codec, état PLC, LR, MLBS).

7.1.2 Impact du choix du codec et du taux de paquet perdu

Dans cette section, nous allons comparer les performances des codecs iLBC, Speex, Silk. Nous avons encodé les mêmes séquences vocales originales avec les trois codecs et dans les mêmes conditions (même fréquence d'échantillonnage, même débit, PLC activé,…), et nous avons appliqué plusieurs processus de perte de paquets pour voir la robustesse de ces codecs.

La Figure 7.1 montre la variation de la qualité des séquences vocales altérées par rapport au taux de paquet perdu (taille moyenne des paquets consécutifs perdus égale à 1). Nous rappelons que la qualité des séquences vocales a été mesurée en utilisant l'outil PESQ, sur une échelle de 0 à 5. Avec un taux de perte de paquets égale à 1%, le codec Silk offre une meilleure qualité (PESQ-MOS ≈ 4) que le codec iLBC (PESQ-MOS ≈ 3,6) et le codec Speex (PESQ-MOS ≈ 3,8). Nous remarquons aussi que, pour un faible taux de perte (entre 1% et 7% de paquet perdu), le codec Silk dépasse, en terme de PESQ-MOS, les codecs iLBC et Speex. A partir de 7% de perte, les codecs Silk et iLBC fournissent une qualité semblable

des séquences vocales. Le codec Speex fourni, dans la majorité des cas, des performances inférieures à celles de Silk et de iLBC.

Nous remarquons aussi, à partir de la Figure 7.1, l'impact que peut avoir les pertes de paquet sur la qualité de la voix. La qualité reste acceptable (PESQ-MOS > 3) pour des taux de perte inférieurs à 10%. A partir de 10% de perte, la qualité de la voix devient mauvaise (PESQ-MOS < 3), et gênante pour les interlocuteurs.

La Figure 7.2, comme la Figure 7.1, montre la robustesse des codecs vis-à-vis des pertes de paquet, dans le cas où la taille moyenne des paquets consécutifs perdus égale à 5. Nous remarquons que les codecs fournissent une qualité équivalente. Silk dépasse légèrement les deux autres codecs. Nous remarquons aussi que la qualité est réduite quand les pertes sont consécutives (MLBS=5) pour un même pourcentage de perte.

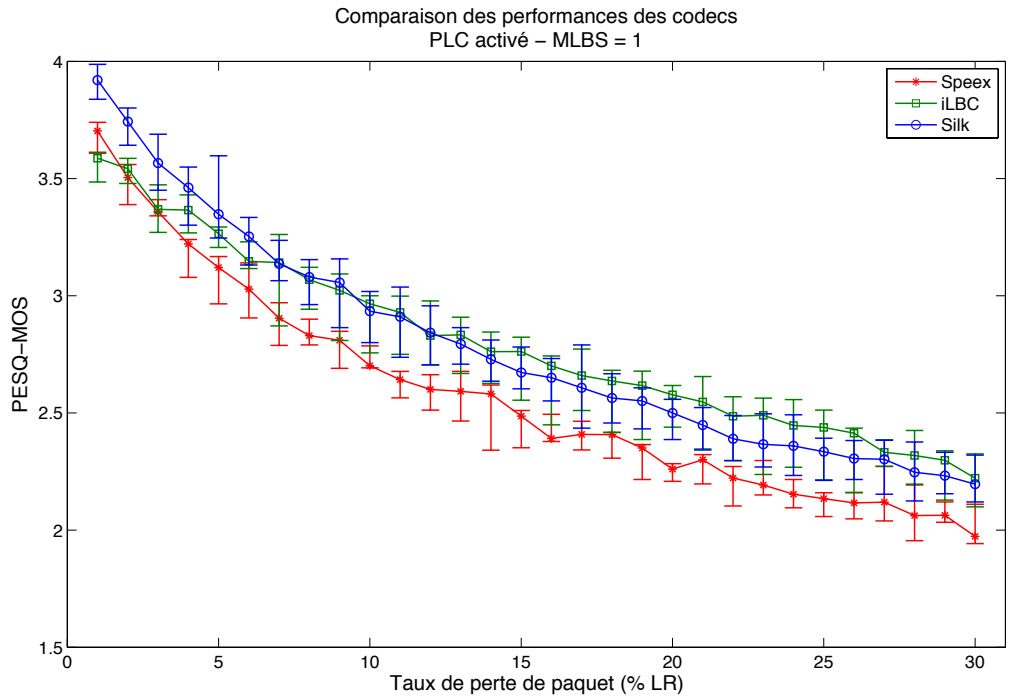


Figure 7.1. Performance des codecs Speex, iLBC et Silk – MLBS=1

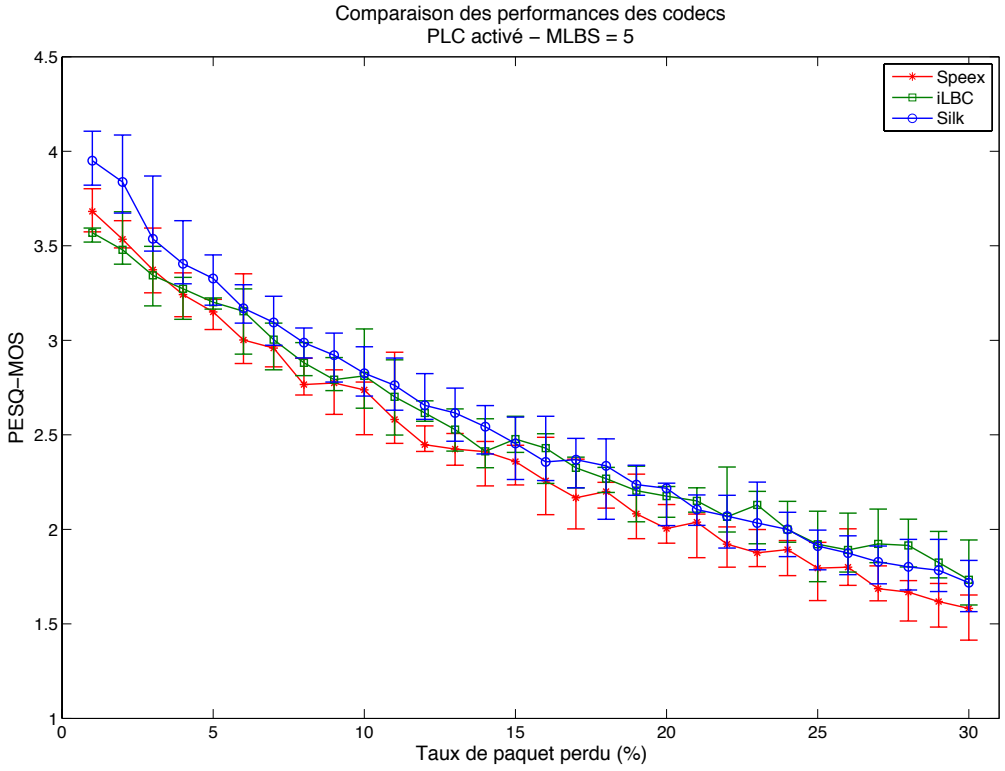


Figure 7.2. Performance des codecs Speex, iLBC et Silk – MLBS=5

7.1.3 Impact du masquage de perte sur la qualité

Dans cette section, nous étudions l'impact que peut avoir le masquage de perte (*Packet Loss Concealement*) sur la qualité des séquences vocales. La Figure 7.3 et la Figure 7.4 montrent l'impact que peut avoir l'activation du PLC dans les cas MLBS=1 et MLBS=5. Dans le cas où le MLBS=1 (Figure 7.3), nous remarquons que l'algorithme de masquage de perte du codec iLBC est plus performant que ceux de Speex et de Silk. En fait, la qualité peut s'améliorer jusqu'à une unité dans le cas du codec iLBC, par exemple, dans le cas où le taux de perte est égale à 25%, la qualité de la voix a été améliorée de PESQ-MOS=1,6 à PESQ-MOS=2,6, grâce à l'algorithme de masquage de perte. Dans le cas où le MLBS=5 (Figure 7.4), les codecs et leurs algorithmes de masquage de perte fournissent des performances équivalentes. Les codecs ont, en général, du mal à corriger les pertes quand il y a beaucoup de paquets consécutifs perdus. L'amélioration de la qualité des séquences vocales, codées avec les trois codecs, atteint ΔPESQ-MOS=+0,7. Les performances de Silk, dans le cas où le PLC n'est pas activé, dépassent celles des codecs Speex et iLBC, dans les deux cas où MLBS=1 et MLBS=5.

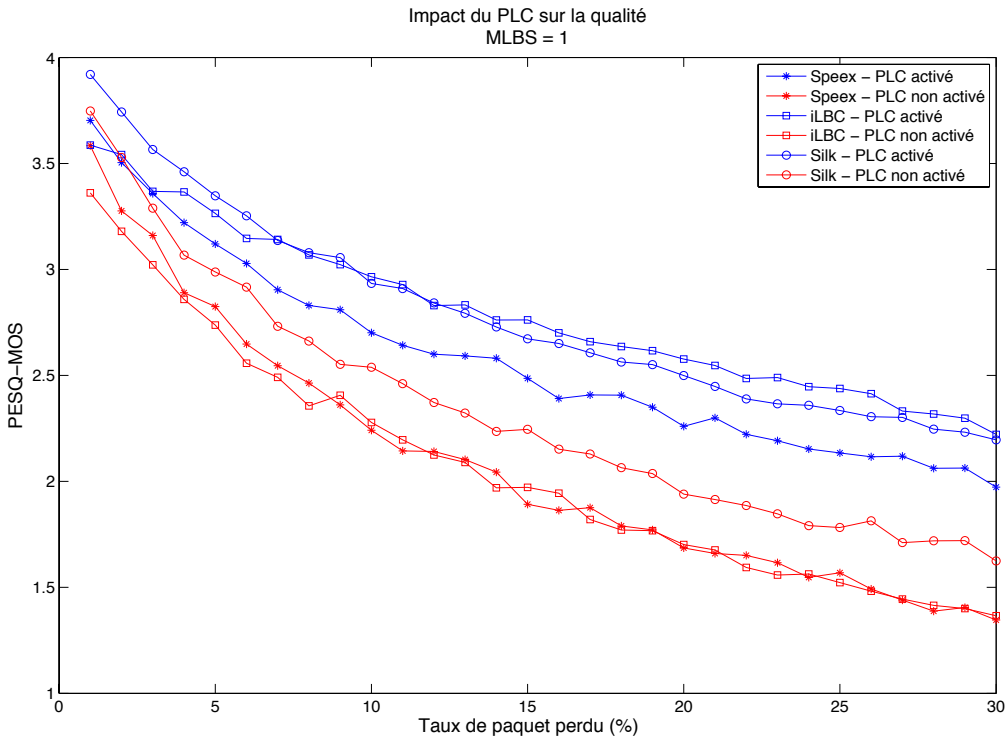


Figure 7.3. Impact du PLC sur la qualité – Cas où MLBS=1

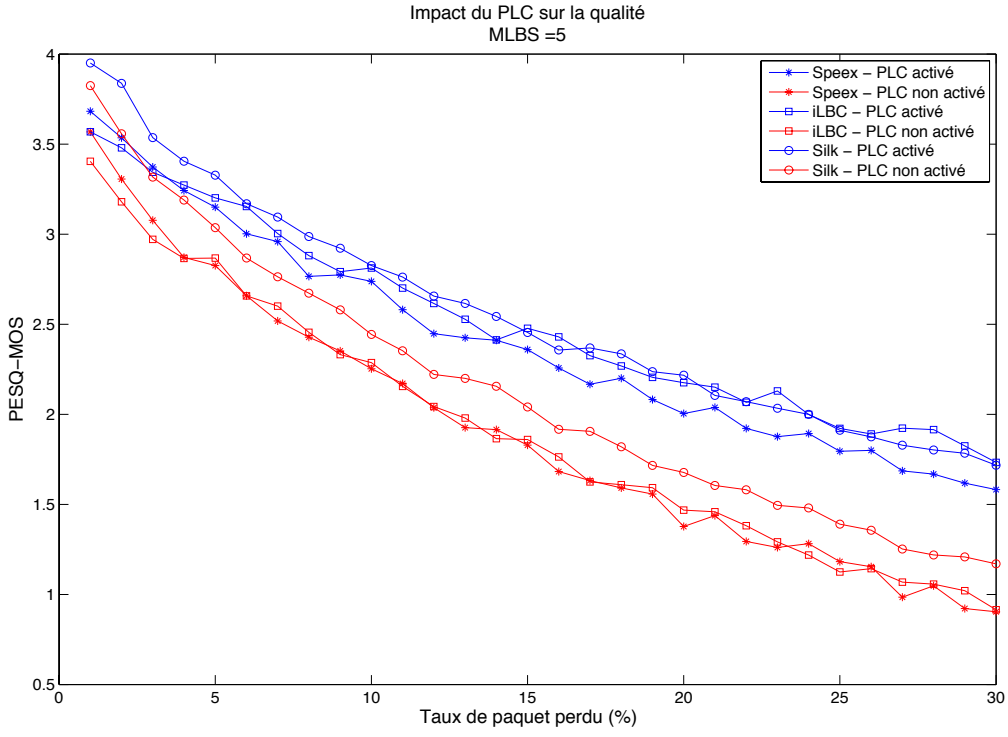
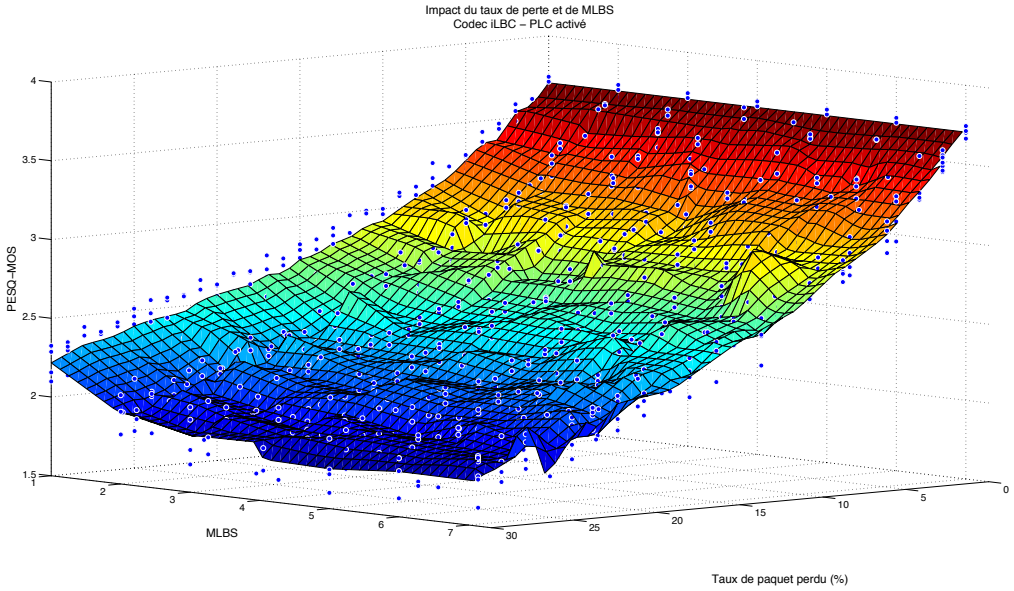


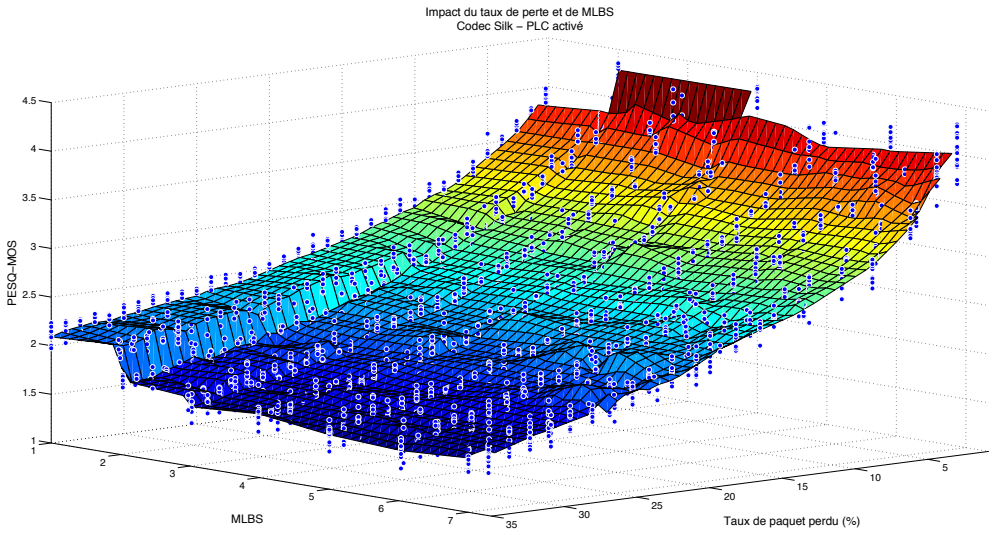
Figure 7.4. Impact du PLC sur la qualité – Cas où MLBS=5

7.1.4 Impact des taux de perte de paquets et du MLBS

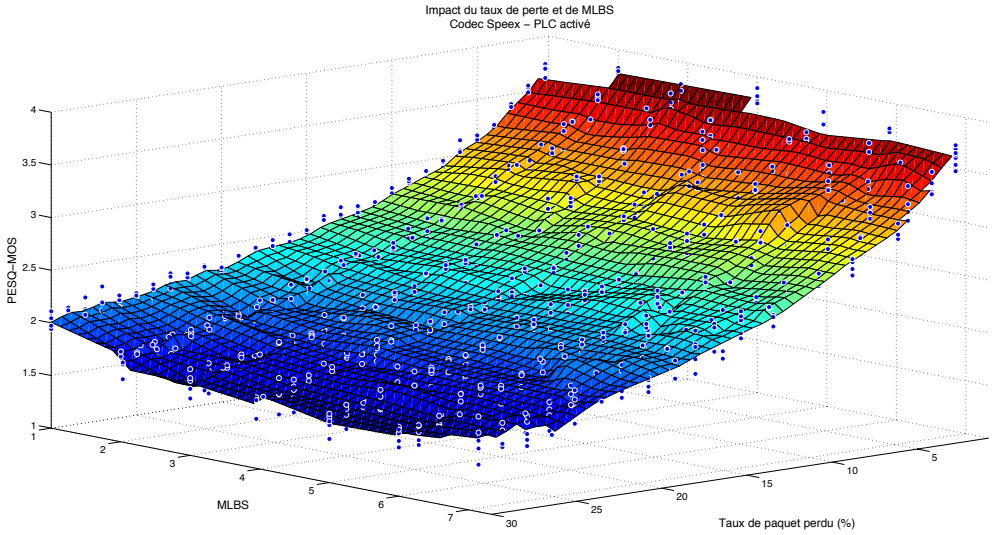
La Figure 7.5 montre la décroissance de la qualité des séquences altérées en fonction des taux de perte de paquets et de la taille moyenne des pertes consécutives. Dans les trois cas (codec iLBC, Speex et Silk), la qualité commence à se dégrader à partir d'un taux de perte supérieur à 10%. Nous remarquons aussi que MLBS peut avoir un impact différent selon le taux de perte. Par exemple, dans le cas d'un faible taux de perte (inférieur à 5%), le facteur MLBS n'a pas d'impact important sur la qualité : les qualités sont presque égaux pour tous les cas de MLBS. Par contre, nous voyons bien l'impact que peut avoir le paramètre MLBS dans le cas où il y a un grand taux de paquet perdus. Nous notons aussi que plus la moyenne des paquets consécutifs perdus est grande, plus la qualité se détériore. En effet, la perte d'un paquet est équivalent à la perte de 20 ms de la conversation, ce qui ne peut ne pas être remarquable et donc non gênant. Par contre, dans le cas de perte consécutives de 20 ms, la conversation risque d'être interrompue durant 140 ms et gênante pour les interlocuteurs.



(a) Codec iLBC



(b) Silk



(c) Speex

Figure 7.5. Variation de PESQ-MOS en fonction du taux de perte et du MLBS (PLC activé)

7.2 Estimation de la qualité

Dans cette section, nous allons étudier différentes techniques pour l'estimation de la qualité à partir des paramètres réseaux et codecs. La première partie détaille la technique de la régression polynomiale, et la seconde partie présente la technique des réseaux de neurones.

7.2.1 Régression polynomiale

Le but de la régression polynomiale est d'ajuster un polynôme sur une série de points expérimentaux, afin de décrire ces points expérimentaux par une loi empirique facile à utiliser.

L'estimation polynomiale est un outil utile pour représenter un ensemble de données de manière linéaire ou quadratique. MATLAB a des fonctions (Polyfit et Polyval), qui peuvent rapidement et facilement adapter un polynôme à un ensemble de données. La formule générale pour un polynôme est :

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_N x^N \quad \textbf{(7.2)}$$

Polyfit permet de générer les coefficients du polynôme qui peut être utilisé pour approximer une courbe en fonction des données.

Les polynômes obtenus sont sous la forme :

$$f(x,y) = p_{00} + p_{10}x + p_{01}y + \ldots + p_{14}xy^4 + p_{05}y^5 \quad \textbf{(7.3)}$$

avec x le taux de perte de paquets et y la valeur de MLBS.

Le Tableau 7-1 résume les valeurs des coefficients des polynômes qui permettent l'estimation de la qualité pour les trois codecs et les cas où le PLC est activé ou non activé.

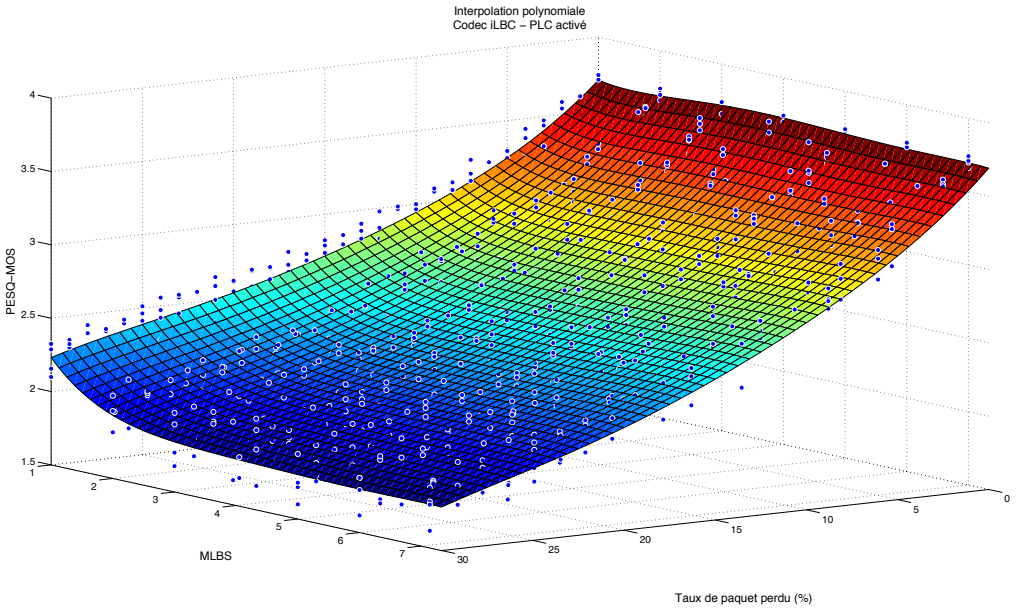
PLC	**PLC non activé**			**PLC activé**		
Codec	**iLBC**	**Speex**	**Silk**	**iLBC**	**Speex**	**Silk**
p_{00}	4,621	4,919	3,759	3,949	-0,2002	3,913
p_{10}	-0,24	-0,2771	-0,3021	-0,08286	0,0973	-0,2375
p_{01}	-1,738	-1,946	0,7619	-0,4524	0,01706	0,4974
p_{20}	0,02562	0,02733	0,03012	0,006027	-0,02699	0,02266
p_{11}	-0,08217	-0,06913	-0,0642	-0,046	0,00747	-0,03831
p_{02}	1,064	1,165	-0,5426	0,2742	-0,000831	-0,2633
p_{30}	-0,001415	-0,001392	-0,001553	-0,0003307	-0,0006783	-0,001182
p_{21}	0,0001996	-0,00108	0,001292	0,0008194	0,0103	0,0008017
p_{12}	0,02913	0,02904	0,01247	0,01158	-0,01523	0,007177
p_{03}	-0,2859	-0,3128	0,1822	-0,07162	2,096e-05	0,0711
p_{40}	3,777e-05	3,647e-05	3,724e-05	9,85e-06	3,243e-05	2,914e-05
p_{31}	4,406e-05	5,346e-05	2,161e-05	-3,334e-06	-2,96e-05	4,685e-06
p_{22}	-0,0003824	-0,0001348	-0,0004998	-0,0001832	-0,001331	-0,0002253
p_{13}	-0,003333	-0,003929	0,000145	-0,001101	0,003083	-0,0001864

p_{04}	0,0349	0,03871	-0,02759	0,008451	-2,094e-07	-0,009387
p_{50}	-3,875e-07	-3,825e-07	-3,357e-07	-1,195e-07	-4,32e-07	-2,711e-07
p_{41}	-6,725e-07	-5,857e-07	-4,689e-07	-6,208e-09	-3,73e-07	-1,428e-07
p_{32}	5,301e-07	-6,013e-07	1,865e-06	7,075e-07	3,349e-06	5,432e-07
p_{23}	2,924e-05	1,33e-05	3,113e-05	1,003e-05	6,133e-05	1,506e-05
p_{14}	0,0001203	0,0001832	-9,905e-05	3,534e-05	-0,0001858	-3,07e-05
p_{05}	-0,001584	-0,001797	0,001524	-0,0003716	-0,2002	0,0004816

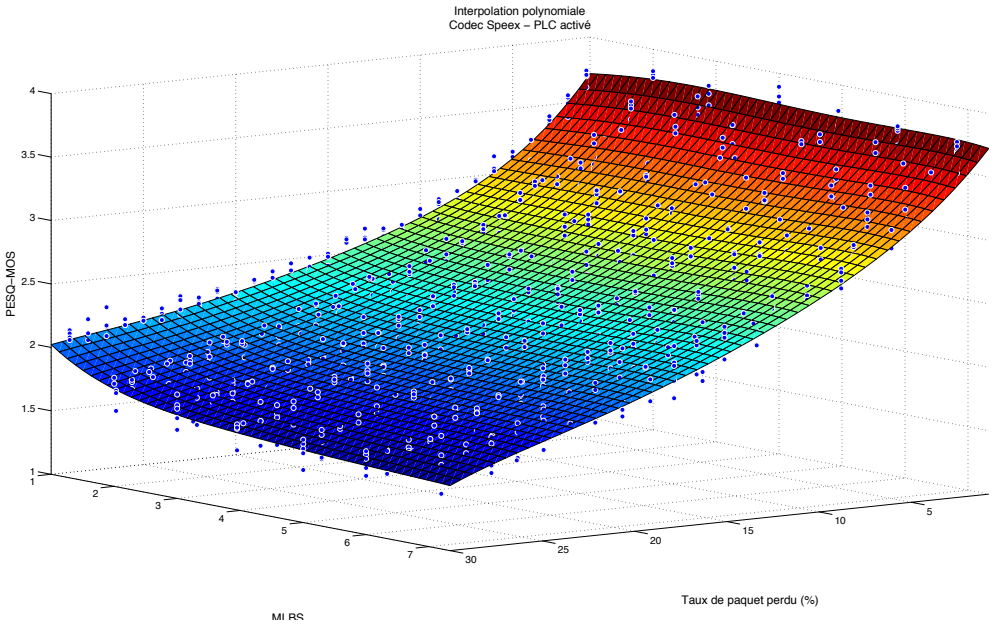
Tableau 7-1. Coefficients des polynômes de régression

La Figure 7.6 montre les interpolations obtenues en utilisant les polynômes pour les trois codecs. Nous remarquons que les surfaces obtenues interpolent bien les scores fournis par PESQ. La qualité se dégrade quand le taux de perte et le MLBS augmentent.

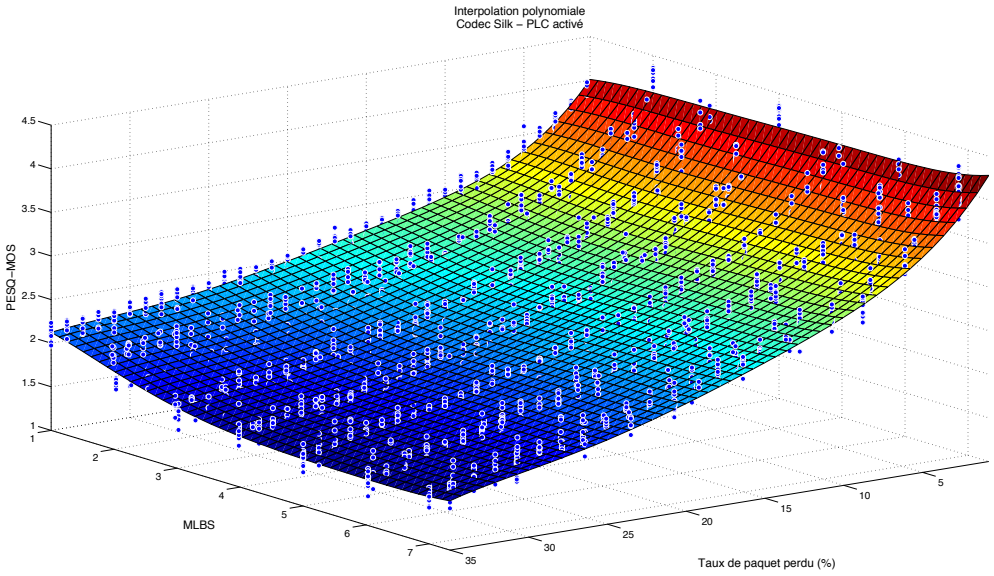
La Figure 7.6 montre les interpolations obtenues en utilisant les polynômes pour les trois codecs. Nous remarquons que les surfaces obtenues interpolent bien les scores fournis par PESQ. La qualité se dégrade quand le taux de perte et le MLBS augmentent.



(a) Codec iLBC



(b) Codec Speex



(c) Codec Silk

Figure 7.6. Interpolation polynomiale

La Figure 7.7 montre la bonne corrélation entre les scores fournis par PESQ et les scores fournis par les polynômes, dans tous les cas possibles (codec, PLC). Le Tableau 7-2 résume les performances des interpolations et confirme l'efficacité de la régression polynômiale pour l'estimation de la qualité. Nous remarquons que tous les polynômes obtenus fournissent une corrélation R > 0,98, ce qui reflète l'efficacité des résultats obtenus. Nous rappelons que plus le facteur de corrélation R est proche de 1, plus nous avons une bonne corrélation.

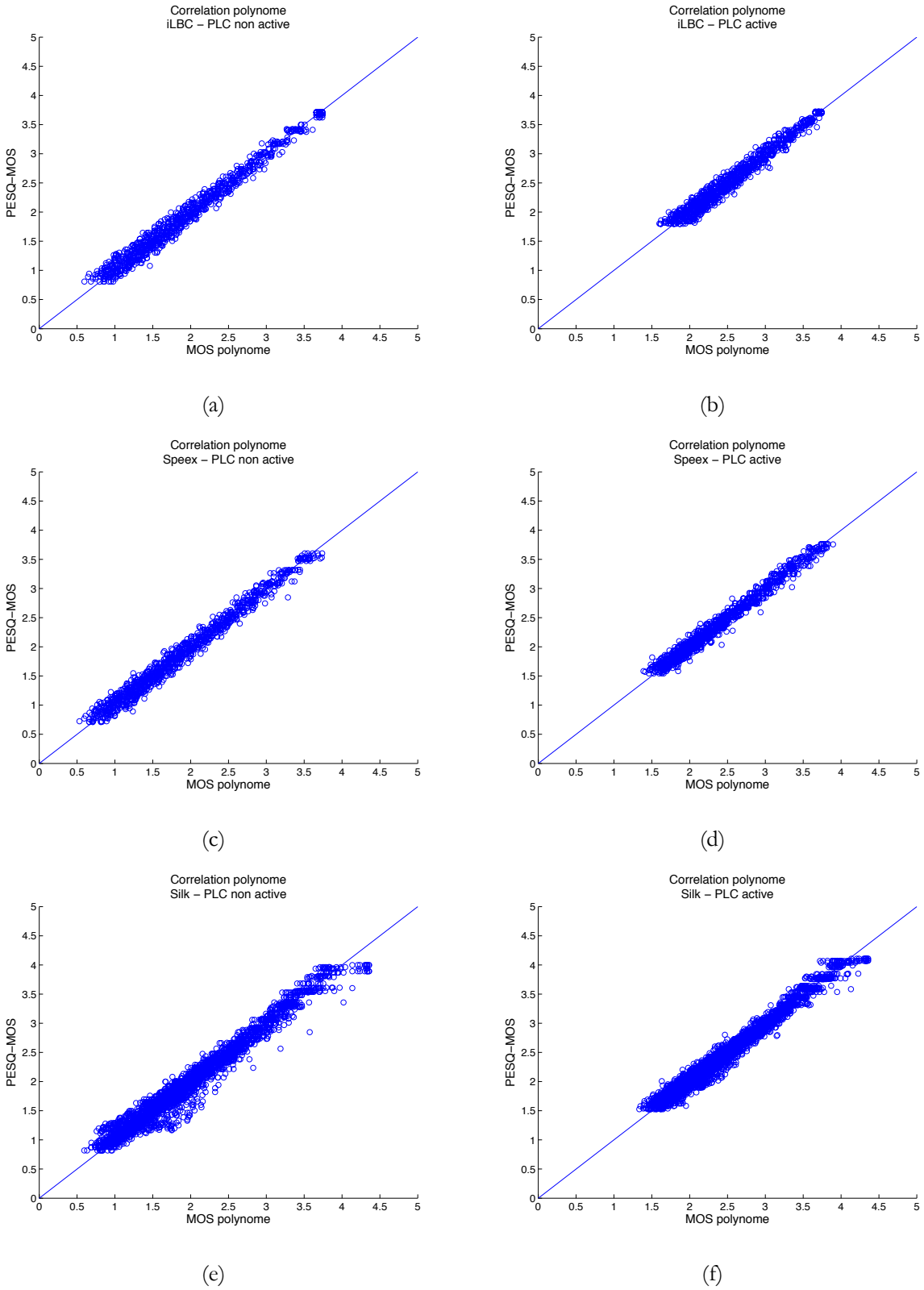


Figure 7.7. Corrélation entre le MOS donné par PESQ et le MOS donné par les polynômes

PLC	PLC non activé			PLC activé		
Codec	**iLBC**	**Speex**	**Silk**	**iLBC**	**Speex**	**Silk**
SSE	9,877	10,97	57,44	8,283	7,86	30,36
R	0,99	0,9905	0,9841	0,9831	0,9886	0,9875
R-square	0,9801	0,9811	0,9686	0,9665	0,9774	0,9753
RMSE	0,108	0,1035	0,1403	0,09889	0,08761	0,102

Tableau 7-2. Performances des polynômes

7.2.2 Estimation avec les Réseaux de Neurones

Dans cette partie, nous allons étudier l'efficacité des réseaux de neurones pour l'estimation de la qualité des séquences vocales. Pour cela, nous avons utilisé des réseaux de neurones à 3 couches comme l'indique la Figure 7.8. Pour ne pas compliquer les réseaux de neurones, nous avons décidé de générer un réseau de neurones pour chaque cas (codec, PLC). De cette façon nous obtenons six réseaux de neurones, avec deux entrés chacun, au lieu d'un seul avec quatre entrés. La première couche du réseau de neurones est composée de deux neurones, qui correspondent respectivement au taux de perte et au MLBS. La couche du milieu (couche cachée) est composée de six neurones. La dernière couche contient un seul neurone qui correspond au résultat final. Le résultat final correspond à une estimation de la qualité.

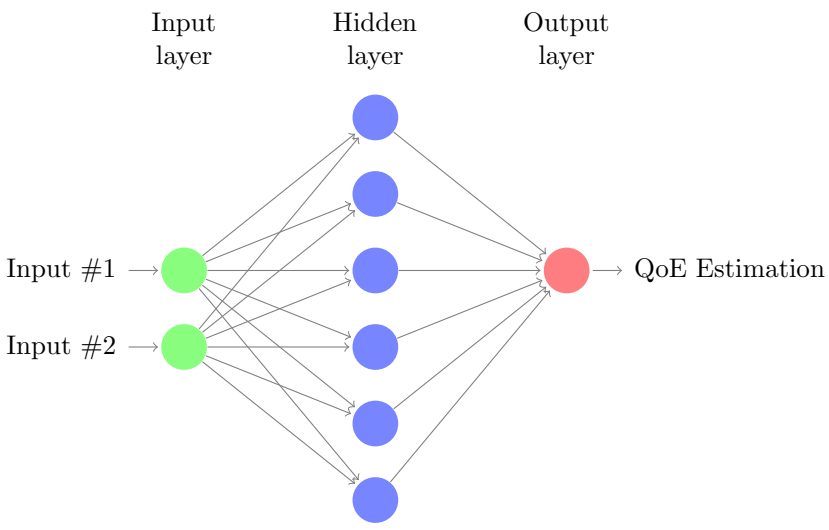
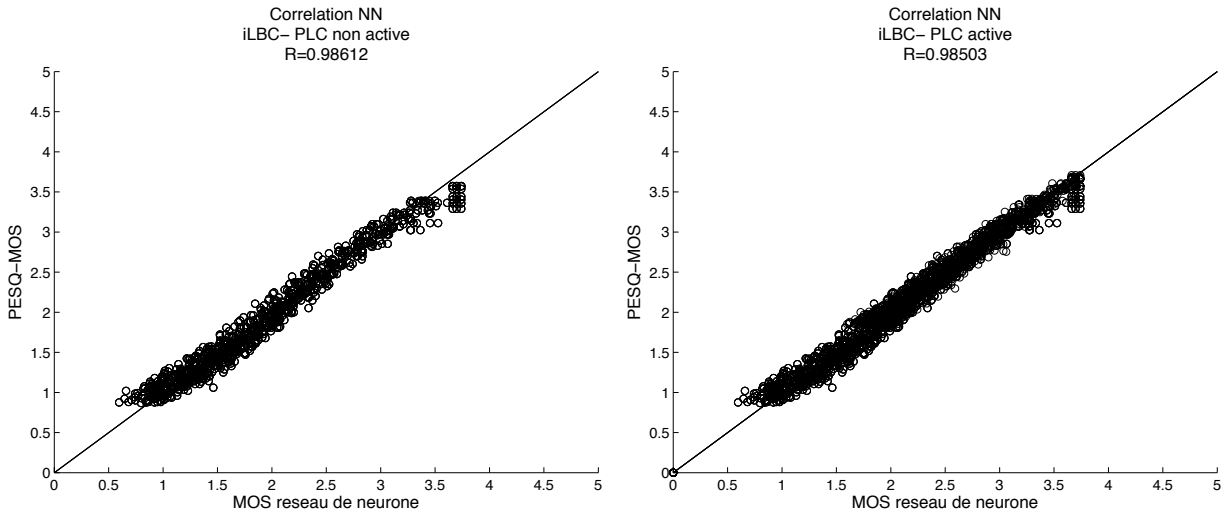


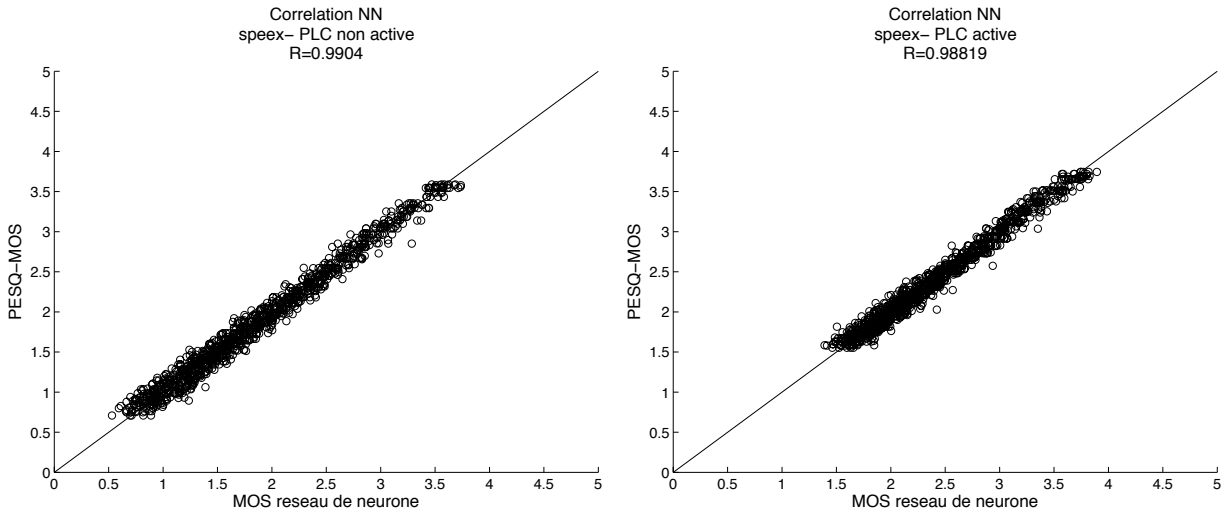
Figure 7.8. Architecture du réseau de neurones

Pour l'obtention d'un réseau de neurones efficace, nous avons utilisé 70% des données pour l'entrainement, 15% pour la validation et 15% pour le test. La Figure 7.9 et le Tableau 7-3 montrent les performances de chacun des réseaux de neurones que nous avons obtenues. Les figures montrent que l'estimation de la qualité en utilisant les réseaux de neurones est très proche de celle fourni par PESQ. Les

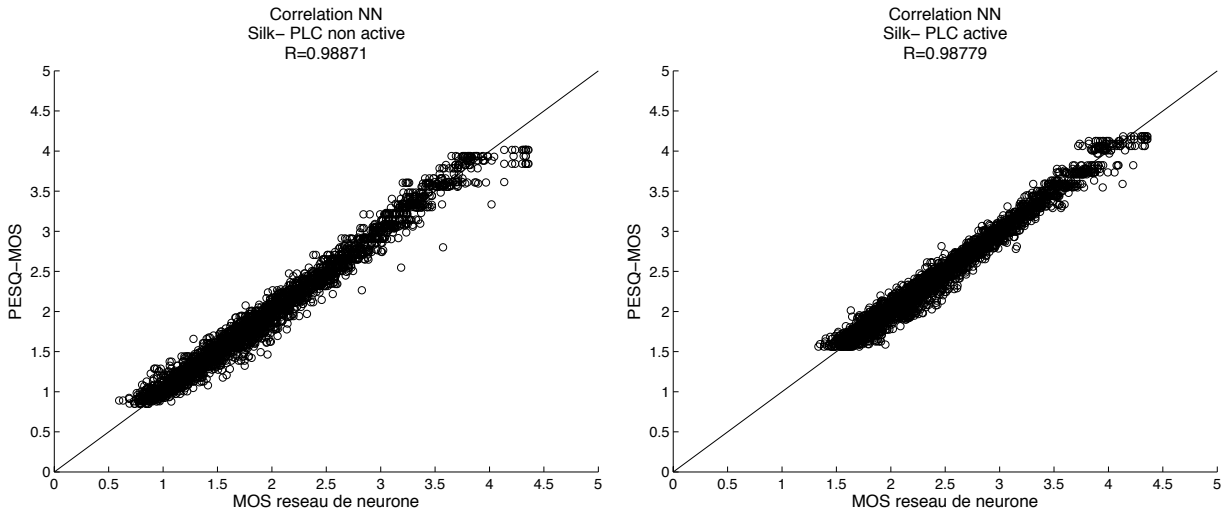
facteurs de mesure de corrélation prouvent aussi qu'il y a une forte corrélation entre les scores de PESQ et ceux des réseaux de neurones.



(a) iLBC



(b) Speex



(c) Silk

Figure 7.9. Corrélation des réseaux de neurones avec les MOS donné par PESQ

PLC	PLC non activé			PLC activé		
Codec	iLBC	Speex	Silk	iLBC	Speex	Silk
SSE	14,2716	11,0773	41,0328	8,4864	8,2099	29,8996
R	0,9861	0,9904	0,9887	0,9850	0,9882	0,9878
R-square	0,9724	0,9809	0,9775	0,9703	0,9765	0,9757
RMSE	0,1282	0,1030	0,1181	0,0989	0,0886	0,1008

Tableau 7-3. Performances des réseaux de neurones

Les résultats obtenus par les réseaux de neurones sont légèrement meilleurs que ceux obtenus par les polynômes. L'avantage de l'utilisation des réseaux de neurones est la possibilité d'utiliser un nombre différent de paramètres et ne pas se limiter à deux comme dans le cas de l'interpolation polynomiale. Ceci est dû au fait que l'interpolation polynomiale est très difficile quand le nombre de paramètres dépasse deux.

7.2.3 Comparaison avec d'autres méthodes

Dans cette section, nous comparons les performances de notre méthode avec deux autres méthodes non intrusives : IQX et l'E-Model. Le codec commun supporté par ces trois méthodes est iLBC, donc la comparaison ne comprendra que les résultats correspondant au codec iLBC.

Comme mentionné dans [79] et [80], les paramètres du modèle E correspondant au codec iLBC sont les suivants :

$$R = 93{,}2 - I_d - I_e - A;$$

$$I_e = 10 + 19{,}8 * log(1 + 29{,}7 * \frac{ploss}{100});$$

$$I_d=0;$$

$$A=0;$$

(7.4)

Nous utilisons les équations (3.12) et (7.4) pour calculer le MOS à partir du E-Model et l'équation (3.14) pour la méthode IQX. Figure 7.10 montre la corrélation entre les prévisions de la qualité vocale donnée par la formule IQX, le E-model et PESQ-PSQA. Nous pouvons remarquer que l'estimation de la qualité par E-Model est globalement bonne pour une qualité vocale moyenne, mais il y a un petit écart quand il n'y a pas de distorsion majeure de la séquence vocale. En outre, la prévision du modèle IQX ne corrèle pas très bien avec PESQ dans le cas d'un taux de perte de paquets faible. Ceci est dû à l'usage de la formule statique dans IQX et l'ignorance de certains facteurs majeurs. De plus, les mauvais résultats du E-Model et de l'IQX peuvent s'expliquer par le fait qu'ils ne prennent pas en considération les pertes par rafales (les pertes consécutives de paquet).

On peut observer que PESQ-PSQA a une bonne corrélation avec les scores de PESQ pour une variété de séquences de parole. PESQ-PSQA fournit également de bons résultats quand il y a une grande quantité de pertes. La méthode proposée est robuste vis-à-vis des pertes en rafales des paquets et compatible avec beaucoup de codec.

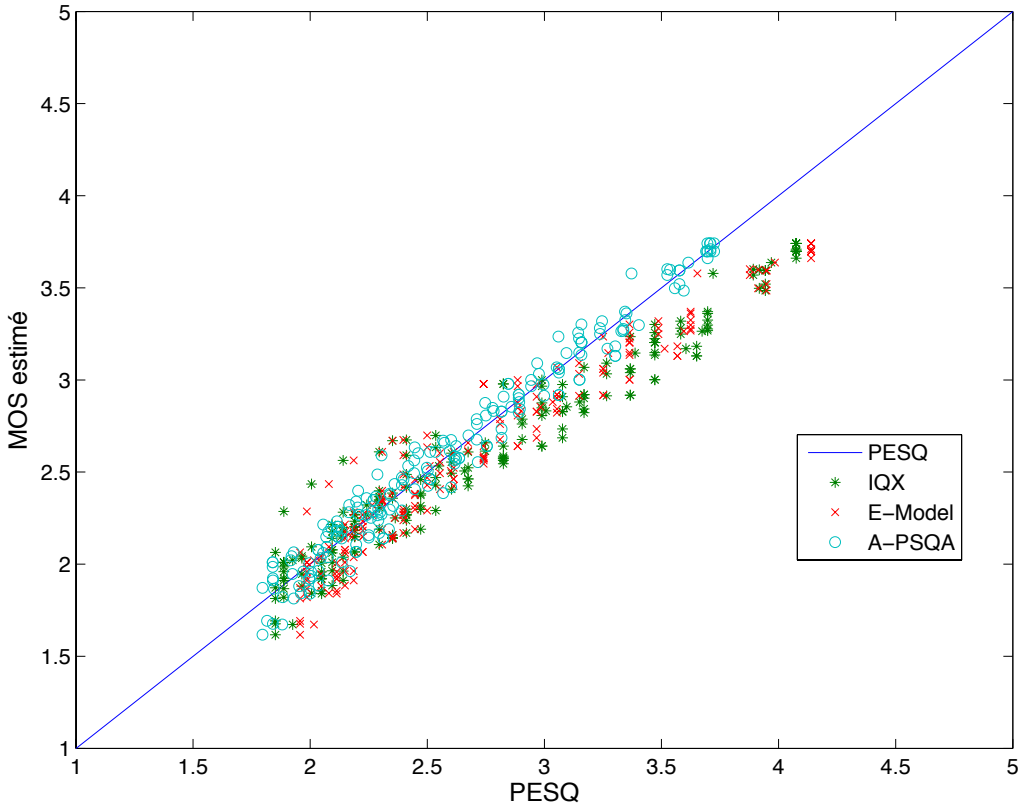


Figure 7.10. Comparaison de la corrélation de la méthode proposée avec les autres méthodes

Conclusion

Dans ce chapitre, nous avons étudié comment les différents paramètres réseaux et applicatifs peuvent affecter la qualité des flux VoIP (à sens-unique, c'est-à-dire, non interactif). Dans la première section, nous avons analysé la qualité vocale perçue comme une fonction de l'ensemble des quatre paramètres considérés (codec, état PLC, taux de perte, MLBS), et nous avons observé comment les différents paramètres interagissent avec la qualité. Nous avons utilisé l'outil PESQ pour avoir une estimation objective de la qualité des séquences vocales. Nous avons remarqué que les paramètres choisis ont un impact relativement différent sur la qualité de la voix.

Dans la deuxième partie de ce chapitre, nous nous sommes intéressés à l'utilisation des réseaux de neurones et de la régression polynomiale dans le contexte de l'estimation de la qualité des séquences vocales. Nous avons fourni deux méthodes qui permettent d'estimer la qualité fournie par PESQ, à partir des paramètres réseaux et applicatifs. L'étude de performance de ces méthodes montre qu'ils fournissent une bonne corrélation avec PESQ. En effet, nous obtenons un facteur de corrélation $R > 0,98$ pour toutes les méthodes proposées (Réseau de neurones et Régression polynomiale).

Ces méthodes obtenues prennent l'avantage de PESQ, en termes de forte corrélation avec le score subjectif, et elles peuvent être utilisées en mode temps réel.

Chapitre 8

8 Mise en œuvre des méthodes proposées de monitoring de QoE dans un terminal utilisateur

Les services multimédias d'aujourd'hui sont de plus en plus dominants dans le monde d'Internet et constituent une source de revenus majeure pour les acteurs concernés. Ainsi, la nécessité de surmonter les faiblesses et les limites de l'Internet actuelle, en matière de communication des médias, devient de plus en plus impérative.

En réponse à ce besoin, le projet ALICANTE propose une approche évolutive conceptuelle et architecturale, orientée vers le déploiement d'un « écosystème des médias ». Il préconise de fournir un environnement flexible dans lequel participe et interagit des services multiples, des fournisseurs de contenu (*Server Providers* SP/*Content Providers* CP), des opérateurs réseaux et les utilisateurs finaux (*End-Users* EU). La solution proposée permet aux utilisateurs finaux d'accéder aux services médias offerts dans des contextes variés, dans un environnement entièrement géré qui permet l'optimisation de la QoS et la QoE, et aussi pour partager et diffuser leurs propres contenus médias de façon dynamique et en toute transparence. De ce fait, il y a un besoin essentiel d'un mécanisme, côté client, qui permet l'estimation de la QoE et le monitoring (surveillance).

Les buts de la surveillance de la QoE dans ce projet, réalisée au sein de l'environnement de l'utilisateur, sont comme suit :

- Premièrement, il permet d'estimer l'expérience utilisateur évaluée, en temps réel, et informer le SP/CP. De cette façon, le SP/CP peut avoir à tout moment une indication sur la qualité du support au niveau de l'utilisateur.
- Deuxièmement, il fournit des informations sur le terminal et le contexte de l'utilisateur au SP/CP (via le profil de l'utilisateur) afin de déclencher une adaptation du flux média pour correspondre à la condition et la capacité du terminal.
- Troisièmement, il fournit des informations sur le terminal et le contexte de l'utilisateur à la Home-Box, afin de déclencher l'adaptation, au niveau de la Home-Box, des flux multimédia transmis.

Les fonctionnalités requises du sous-système QoE Monitoring peuvent être classées comme suit :

- Suivi des sessions multimédias présentés dans le terminal de l'utilisateur final, en se concentrant sur les paramètres au niveau du réseau (débit, perte de paquets, la gigue, la duplication, etc.) ;
- Évaluation de la qualité de l'expérience, telle qu'elle est perçue par l'utilisateur final ;

- Surveillance de l'état des terminaux et des ressources ;
- Communication de tous les paramètres du gestionnaire de profils utilisateur.

Le sous-système de surveillance QoE, au niveau du terminal de l'utilisateur final, est divisé en trois sous-modules, chacun d'eux entreprend une tâche différente :

- un outil de suivi de la qualité de service dans l'environnement utilisateur (*User Environment* UE), noté *QoE Monitoring Tool @ UE*, qui surveille tous les flux médias entrants et effectue une analyse de trafic ;
- un évaluateur de QoE, noté *QoE Evaluator*, qui reçoit les paramètres réseau et calcule une estimation de la QoE en utilisant sur les réseaux de neurones ;
- un gestionnaire de surveillance QoE, noté *QoE Monitoring Manager*, qui rassemble les ressources du terminal, reçoit les paramètres issus de *QoS Monitoring Tool @ UE* et de *QoE Evaluator*, et envoie tous ces paramètres au profil de l'utilisateur, noté *User Profil*, via le manager du profil de l'utilisateur, noté *User Profile Manager*.

8.1 QoE Monitoring Manager

Le module *QoE Monitoring Manager* est le module de coordination du sous-système de surveillance de la QoE. La Figure 8.1 montre le sous-système de surveillance de la QoE et les modules qui interagissent.

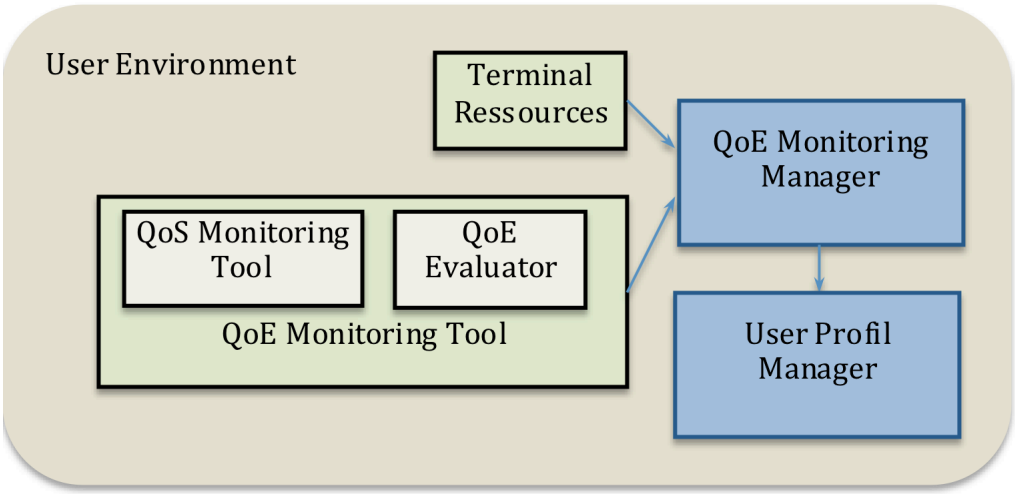


Figure 8.1. Schéma du sous-système de monitoring de la QoE

Le module *QoE Monitoring Manager* est un logiciel installé sur le terminal utilisateur, fonctionnant en permanence en arrière-plan, totalement transparent pour l'utilisateur.

Le module *QoE Monitoring Manager* est en interfaçage avec :

- le module *QoS Monitoring Tool @ UE*, pour extraire des métriques au niveau réseau, session et média des sessions multimédias présentées au terminal ;

- le module *QoE Evaluator*, pour récupérer une estimation du score MOS pour les sessions multimédias présentées au terminal ;
- le système d'exploitation du terminal, pour récupérer les paramètres de fonctionnement tels que la charge CPU, état de la batterie, etc ;
- le module *User Profile Manager* (situé à la Home-Box et gère la base de données de la Home-Box HBDB) pour stocker et mettre à jour les paramètres dérivés, au niveau la partie dynamique du profil utilisateur.

8.2 QoS Monitoring Tool @ UE

Le module *QoS Monitoring Tool @ UE* analyse en temps réel le trafic multimédia reçu par le terminal et extrait des paramètres objectifs (métriques réseau, métriques session, métriques média). La capture (*sniffing*) du flux de données agit au niveau de la couche liaison de données, et donc n'interfère pas avec le lecteur multimédia utilisé par l'utilisateur pour la lecture du contenu des médias. Le décodage audio et vidéo est partiellement effectué, vu que la reconstruction complète des échantillons audio/vidéo n'est pas nécessaire pour évaluer les paramètres de qualité; principalement, l'analyse est effectuée au niveau des en-têtes, par l'extraction des paramètres pertinents, tels que le paramètre de quantification vidéo.

La Figure 8.1 montre le schéma général de *QoE Monitoring Tool*, comme une combinaison du module *QoS Monitoring Tool @ UE* et du module *QoE Evaluator*.

QoE Monitoring Tool est le résultat de la combinaison des deux sous-entités suivantes :

- *QoS Monitoring Tool @ UE* : ce module effectue la capture (*sniffing*) du flux de données en temps réel et calcul les paramètres objectives QoS ;
- *QoE Evaluator* : ce module prend en entrée les paramètres de qualité de service fournis par le *QoS Monitoring Tool @ UE* et calcul le score MOS. Le calcul de la QoE est basé sur les réseaux de neurones.

8.3 QoE Evaluator

Le module *QoE Evaluator* effectue l'estimation du MOS, en utilisant le réseau de neurones (PSQA) (approprié au service/codec), en se basant sur les paramètres objectifs de qualité de service fournis par *QoS Monitoring Tool @ UE*.

Le *QoE Evaluator* permet donc d'estimer la qualité d'un service (Vidéo sur IP ou Voix sur IP) perçue par l'utilisateur final. En effet, le *QoE Evaluator* est l'ensemble des réseaux de neurones que nous avons proposés dans les chapitres précédents. Les réseaux de neurones utilisent les paramètres du réseau et du codec afin de fournir une estimation de la QoE.

L'estimation de la QoE est ensuite envoyée au module *QoE Monitoring Manager*, pour mettre à jour le profil de l'utilisateur et entreprendre l'adaptation adéquate, si nécessaire, pour améliorer la satisfaction de

l'utilisateur. La Figure 8.2 résume les messages échangés et les interfaces entre les modules du *QoE Monitoring Tool*.

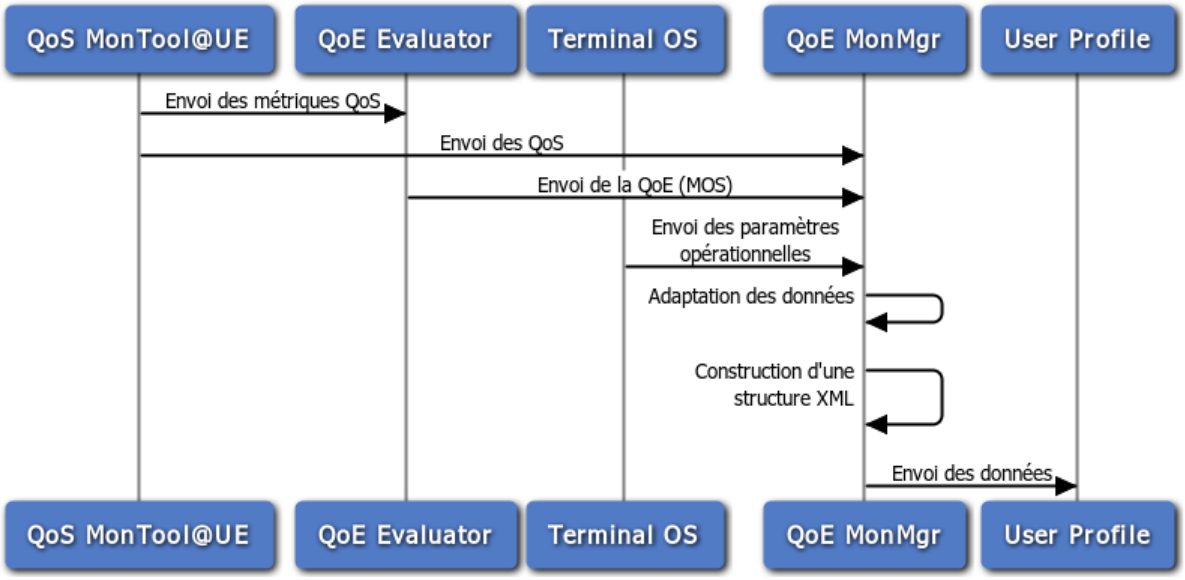


Figure 8.2. Diagramme de séquence du mode opératoire du module QoE Monitoring Tool

Conclusion

Ce chapitre présente la spécification du sous-système de surveillance de la QoE, dans le cadre de l'environnement utilisateur du projet ALICANTE. Le chapitre expose les fonctionnalités des trois sous-modules (*QoE Monitoring Manager*, *QoS Monitoring Tool @ UE*, *QoE Evaluator*).

Grâce à ce sous-système de monitoring et d'estimation de la QoE, les fournisseurs de service et de réseau peuvent mettre en place les actions nécessaires pour fournir au client final une qualité d'expérience optimale. Les actions d'adaptation (comme l'augmentation/diminution du débit binaire, suppression/ajout des couches d'amélioration SVC,…) doivent prendre en compte les capacités du terminal, du service proposé,…

Chapitre 9

9 Conclusion générale et perspective

L'évaluation de la qualité d'expérience des services médias est un problème majeur dans le contexte de l'Internet du Futur. Il existe plusieurs méthodes dans la littérature qui utilisent soit des paramètres réseaux, soit des caractéristiques du signal voix ou vidéo. Ces méthodes ont plusieurs limitations, telles que l'impossibilité de les utiliser dans un contexte temps-réel, la complexité de leurs algorithmes, ou même leurs précisions comparées aux tests subjectifs. Ces limitations nous ont poussés à mettre en place de nouvelles techniques pour estimer la qualité d'expérience.

Dans cette thèse, nous avons proposé plusieurs contributions dans le domaine de l'évaluation de la qualité des services multimédias. Nous nous sommes basés sur la méthodologie PSQA (*Pseudo-Subjective Quality Assessment*) pour mieux étudier la variation de la qualité des flux multimédias transmis sur un réseau tel que l'Internet et pour estimer en temps réel cette qualité. Nous nous sommes intéressés à deux applications multimédia qui ont gagné en popularité au cours des dernières années : la Voix sur IP et la Vidéo sur IP.

9.1 Résumé des contributions

Notre première contribution a consisté en une étude approfondie de la performance de PSQA (dans le cas du codec MPEG-2), qui vise (1) à tester sa robustesse et (2) à la comparer avec d'autres méthodes d'évaluation de la qualité. Les résultats de cette étude sont présentés dans la section 4.2, et montrent que PSQA fournit des estimations du MOS très proches des tests subjectifs, contrairement aux autres méthodes.

Notre deuxième contribution est une sorte de recommandation d'encodage SVC. En effet, nous avons évalué les performances des différents encodeurs SVC dans différentes conditions. Nous avons étudié l'impact de certains paramètres du codage SVC sur la qualité, le débit binaire et le temps d'encodage. Les résultats, dans le chapitre 5, fournissent une recommandation sur les débits binaires à utiliser dans le cas d'un encodage vidéo en haute définition, qui permet d'avoir une très bonne qualité. Les résultats fournissent aussi un comparatif sur la qualité et le débit binaire que nous pouvons obtenir quand nous varions le facteur de quantification QP entre les couches d'amélioration. Nous avons étudié aussi dans cette contribution l'effet que peuvent avoir le mode CGS et le mode MGS sur la qualité des vidéos.

La troisième contribution est la proposition d'un outil d'évaluation de la qualité des vidéos SVC. Dans cette contribution, nous nous sommes basés sur les résultats d'une évaluation objective (en utilisant VQM), pour entrainer un réseau de neurones la relation qui existe entre (i) les paramètres réseaux/vidéos et (ii) la qualité de la vidéo. Le choix de VQM comme outil d'évaluation objectif est dû au fait qu'il corrèle bien avec les scores subjectifs. En utilisant cet outil, nous avons pu évaluer la qualité d'un grand nombre

de vidéos d'une façon automatique et sans les désagréments des tests subjectifs. A partir de ces évaluations, nous avons entrainé un réseau de neurones pour estimer la QoE. Les résultats du chapitre 6 montrent que l'outil proposé corrèle bien avec les scores de VQM.

La quatrième contribution de notre travail consiste en une analyse approfondie de la qualité des communications VoIP et de la façon dont elle est affectée par divers facteurs liés à la fois au réseau et à l'application. Nous avons prêté une attention particulière à la performance des différents encodeurs VoIP (iLBC, Speex, Silk) et à leurs systèmes de correction des erreurs, ce qui pourrait être particulièrement utile dans les applications de contrôle de qualité. Les résultats ont montré que les codecs étudiés fournissent des qualités équivalentes, mais ils ont des comportements différents quand il y a des pertes consécutives de paquets ou quand le masquage de perte est activé. Suite aux résultats de cette étude, nous avons mis en place deux méthodes pour évaluer la QoE en temps réel des communications VoIP. La première méthode est basée sur la régression polynomiale, et la seconde méthode est basée sur PSQA. Ces deux méthodes ont donné des résultats très proches des résultats de PESQ (un outil intrusif d'évaluation objective), comme le montre le chapitre 7. Les résultats obtenus, qui montrent comment les différents facteurs considérés affectent la qualité perçue, peuvent ouvrir la voie à l'amélioration de la qualité, par exemple sous la forme de meilleurs mécanismes de contrôle en temps réel pour les applications multimédia.

La dernière contribution est l'intégration de ces outils dans un ensemble de solutions de « Monitoring QoE » afin de satisfaire au mieux l'expérience des utilisateurs. Cela est proposé dans le cadre du projet ICT-ALICANTE.

9.2 Perspectives

Plusieurs extensions aux travaux proposés au sein de cette thèse sont possibles et envisageables :

- **Les tests subjectifs :** Dans nos travaux, nous avons délibérément évité d'utiliser les tests subjectifs pour évaluer la qualité des séquences vocales ou vidéos. Comme cités précédemment, les tests subjectifs sont très durs à réaliser, surtout quand il y a un grand nombre de vidéos à évaluer et que le nombre d'observateurs est limité. Nous envisageons à l'avenir d'utiliser une base de données qui regroupe un grand nombre de tests subjectifs réalisés selon les normes ITU. Ainsi, nous pourrons avoir des résultats plus précis et nous pourrons généraliser notre outil à d'autres codecs/paramètres vidéo.
- **POLQA :** Parmi les outils d'évaluation de la qualité de la voix, il y a le nouvel outil POLQA. Cet outil permet d'évaluer la qualité d'une façon plus précise que PESQ et est applicable pour un plus grand nombre de séquences vocales. Nous n'avons pas pu l'utiliser pour nos tests parce qu'il n'en existe toujours pas de version complète open source ou académique disponible. Néanmoins, dès que cela sera possible, nous envisagerons de nous en servir.
- **Visio-conférence :** La visio-conférence (ou la vidéo-conférence) est le service qui permet d'inclure la vidéo dans une conversion VoIP. Nous avons proposé dans cette thèse des méthodes qui permettent d'estimer la qualité d'expérience dans le cas de la VoIP et de l'IPTV. Nous proposons donc de « combiner » les deux outils afin d'en créer un qui permettra d'évaluer la satisfaction de l'utilisateur final de visio-conférence. Généralement, les utilisateurs des services de visio-conférence sont plus sensibles à la qualité de la voix qu'à la qualité de la vidéo. Il faut donc

faire un compromis entre la qualité de la voix et celle de la vidéo. L'outil pourra également par la suite, si cela s'avère nécessaire, réduire le débit binaire de la vidéo afin de laisser plus de bande passante à la voix.

- **Service Web :** Nous envisageons de travailler sur l'estimation de la qualité d'expérience dans le cas d'applications web, induisant une navigation et un chargement de données sur des pages Internet. En fait, les pages Internet se chargent avec des vitesses différentes selon leur contenu, les caractéristiques de la connexion Internet et les capacités du serveur qui héberge la page Internet demandée. La lenteur d'affichage d'une page Internet peut causer une gêne aux utilisateurs finaux. C'est pour cette raison qu'il faut mettre en place un outil qui permet d'estimer leur satisfaction lors d'utilisation d'applications incluant un chargement de pages web.
- **MPEG-DASH :** *Dynamic Adaptative Streaming over HTTP* est un nouveau standard utilisé pour le streaming vidéo. MPEG-DASH découpe une séquence vidéo en plusieurs segments, enregistrés chacun dans un fichier à part. Chaque segment contient un court intervalle de temps de lecture de la vidéo. Le contenu (la vidéo) est encodé dans différents débits binaires, et le client sélectionne automatiquement le segment en fonction des conditions actuelles du réseau. L'utilisation du MPEG-DASH est basée sur le protocole HTTP, qui utilise le protocole de transport TCP. Avec ces protocoles, il ne risque pas d'y avoir des pertes de paquets, c'est pourquoi nos méthodes d'évaluation ne s'appliqueront pas à ce genre de service. Nous proposerons alors de mettre en place une méthode qui permet d'évaluer la QoE à partir du débit binaire, la résolution et le facteur de quantification.
- **L'adaptation :** Les méthodes que nous avons présentées dans ce document permettent l'évaluation de la QoE. Nous proposons de mettre en place un outil qui permette de prendre des décisions d'adaptation pour améliorer cette QoE. Comme dans le cas de l'estimation de la qualité de la VoIP, cet outil permettra par exemple d'imposer au client d'utiliser le codec qui fournit la meilleure qualité (sous les mêmes conditions réseaux). De cette façon, le client aura toujours la meilleure qualité sans qu'il doive intervienne.

Publications issues de cette thèse

• Conférences internationales avec comité de lecture

[1] W. Chérif, A. Ksentini, D. Négru, M. Sidibé, "**A_PSQA: Efficient real-time video streaming QoE tool in a Future Media Internet context**", IEEE International Conference on Multimedia and Expo - ICME 2011, Barcelone, Spain.

[2] W. Chérif, A. Ksentini, D. Négru, and M. Sidibé, "**A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services**", IEEE International Conference on Communications - ICC 2012, Ottawa, Canada.

[3] W. Chérif, A. Ksentini, D. Négru, "**No-Reference Quality of Experience estimation of H264/SVC stream**", IEEE Workshop on Quality of Experience for Multimedia Communications - QoEMC 2012, in conjunction with IEEE Globecom 2012, California, USA.

[4] M. Grafl, C. Timmerer, H. Hellwagner, W. Chérif, D. Négru, S. Battista, "**Scalable Video Coding Guidelines and Performance Evaluations for Adaptive Media Delivery of High Definition Content**", IEEE Symposium on Computers and Communications - ISCC 2013, Split, Croatia.

[5] M. Grafl, C. Timmerer, H. Hellwagner, W. Chérif, A. Ksentini, "**Evaluation of Hybrid Scalable Video Coding for HTTP-based Adaptive Media Streaming with High-Definition Content**", IEEE WoWMoM Workshop on Video Everywhere - ViDEv 2013, Madrid, Spain.

• Conférence nationale sans comité de lecture

W. Chérif, "**Context Adaptation based on Quality of Experience in Next Generation Network**", Poster à l'école d'été "ResCom 2011 : Thématiques d'avenir dans les réseaux de communication".

Références

[1] ITU-T SG12, "Definition of Quality of Experience", TD 109rev2 (PLEN/12), Geneva, Switzerland, 16-25 Jan 2007.

[2] ITU-T Recommendation T.81, JPEG Standard, JPEG ISO/IEC 10918-1.

[3] D. LeGall, "MPEG: A video compression standard for multimedia applications", Commun. ACM, vol. 34, no. 4, pp.46 -58 1991

[4] A. Puri, "Video coding using the MPEG-2 compression standard", Proc. SPIE Visual Communications and Image Processing, pp.1701 -1713 1993

[5] A. Wong and C.-T. Chen, "A comparison of ISO MPEG1 and MPEG2 video coding standards", Proc. SPIE Visual Communications and Image Processing, pp.1436 -1448 1993

[6] ITU-T Rec. H.264, ISO/IEC 14496-10:2009, "Advanced video coding for generic audiovisual services", Information Technology-Coding of Audio-Visual Objects, Part 10: Advanced Video Coding, 2010.

[7] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," IEEE Trans. on Circuits and Systems for Video Technology, July 2003.

[8] Stephan Wenger. "H.264/AVC over IP", IEEE Transactions on Circuits and Systems for Video Technology, 13(7):645–656, July 2003.

[9] H. Schwartz et al., Fraunhofer HHI: "The Scalable Video Coding Amendment of the H.264/AVC Standard", http://ip.hhi.de/imagecom_G1/savce/

[10] M. Eberhard, L. Celetto, C. Timmerer, E. Quacchio, H. Hellwagner, and F. S. Rovati, "An interoperable streaming framework for Scalable Video Coding based on MPEG-21," pp. 723-728, Aug. 2008.

[11] Z. Avramova, D. De Vleeschauwer, K. Spaey, S. Wittevrongel, H. Bruneel, and C. Blondia, "Comparison of simulcast and scalable video coding in terms of the required capacity in an IPTV network," Proceedings of Packet Video (PV), pp. 113-122, 2007.

[12] T. Schierl, T. Stockhammer, and T. Wiegand, "Mobile Video Transmission Using Scalable Video Coding," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 17, no. 9, pp. 1204-1217, 2007.

[13] P. Lambert, P. Debevere, S. Moens, R. Van de Walle, and J.-F. Macq, "Optimizing IPTV video delivery using SVC spatial scalability," pp. 89-92, May. 2009.

[14] N. Ramzan, E. Quacchio, T. Zgaljic, L. Celetto, E. Izquierdo, and F. S. Rovati, "Peer-to-peer streaming of scalable video in future Internet applications," Communications Magazine, IEEE, vol. 49, no. 3, pp. 128-135, Mar. 2011.

[15] H. Shen, X. Sun, F. Wu, and S. Li, "Scalable Video Adaptation for IPTV," in proc. IPTV services over World Wide Web Workshop WWW2006, 2006.

[16] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", June 2002

[17] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications", September 2008.

[18] H. R. Sheikh, M. F. Sabir, & A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms", Image Processing, IEEE Transactions on 15(11): 3440– 3451. 2006.

[19] NTIA VQM, http://www.its.bldrdoc.gov/vqm

[20] ITU-R Rec. BT.1683, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference", 2004.

[21] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Transactions on broadcasting, vol. 50, no. 3, pp. 312-322, September 2004.

[22] A. Webster et al. "An objective video quality assessment system based on human perception", Proc. SPIE Conference on Human Vision, Visual Processing, and Digital Display IV, Vol. 1913, San Jose, CA, USA, pp. 15–26.

[23] I. Gunawan, M. Ghanbari, "Image quality assessment based on harmonics gain/loss information", Image Processing, 2005. ICIP 2005. IEEE International Conference on, Vol. 1, pp. I–429–32.

[24] L. Ma, S. Li, and K. Ngi Ngan, "Reduced-Reference Video Quality Assessment of Compressed Video Sequences", IEEE Transactions on circuits and systems for video technology, vol. 22, no. 10, october 2012

[25] I. Gunawan, M. Ghanbari, "Efficient reduced-reference video quality meter, Broadcasting", IEEE Transactions on 54(3): 669–679.

[26] B. Shao, et al. "An adaptative system for real-time scalable video streaming with end-to-end QoS control". WIAMIS, 2010.

[27] M. Sidibé, et al. "A novel monitoring architecture for media services adaptation based on network QoS to perceived QoS mapping". SIViP(2), No. 4, December 2008.

[28] Minoli D., Minoli E., "Delivering Voice over IP Networks", Wiley Computer Publishing, 1998

[29] ITU-T Recommendation G.711 (11/88), "Pulse Code Modulation (PCM) of Voice Frequencies", International Telecommunication Union, November 1988.

[30] International Telecommunication Union Std. G.729 (01/2007), "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-CodeExcited Linear Prediction (CS-ACELP)", January 2007.

[31] International Telecommunication Union Std. G.723.1 (05/2006), "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s", May 2006.

[32] Digital cellular telecommunications system (Phase 2+); Full rate speech; Transcoding (GSM 06.10 version 8.1.1), European Telecommunications Standards Institute (ETSI) European Standard (Telecommunications series) ETSI EN 300 961 V8.1.1, November 2000.

[33] Digital cellular telecommunications system (Phase 2); Half rate speech; Part 2: Half rate speech transcoding (GSM 06.20 version 4.3.1), European Telecommunications Standards Institute (ETSI) Std. ETS 300 581-2, May 1998.

[34] AMR speech Codec; General Description (Release 9), 3GPP Technical Specification 3GPP TS 26.071 V9.0.0, December 2009.

[35] Internet Low Bit Rate Codec (iLBC), Global IP Sound Request for Comments RFC 3951, December 2004.

[36] The Speex Codec Manual Version; 1.2 Beta 3, Xiph.org Foundation Std., December 2007.

[37] SILK Speech Codec, Skype Technologies S.A. Internet-Draft, March 2010.

[38] ITU-T Recommendation G.114. "One-way transmission time", 2003.

[39] C. Boutremans, "Delay Aspects in Internet Telephony", PhD thesis, EPFL, Lausanne, Switzerland, 2002.

[40] L. Sun and E. Ifeachor, "Perceived Speech Quality Prediction for Voice over IP-based Networks", In Proceedings of IEEE International Conference on Communications (IEEE ICC'02), pages 2573–2577, New York, USA, April 2002.

[41] J. F. Kurose and K. W. Ross, "Computer Networking A Top-Down Approach Featuring the Internet", Pearson Addison Wesley, 2001. ISBN 0-201-47711-4.

[42] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and Modelling of the Temporal Dependence in Packet Loss," in Proceedings of IEEE INFOCOM 99, vol. 1, (New York, USA), pp. 345–352, March 1999.

[43] W. Jiang and H. Schulzrinne, "QoS Measurement of Internet Real-Time Multimedia Services," Technical Report, CUCS-015-99, Columbia University, Dec. 1999.

[44] H. Sanneck, "Packet Loss Recovery and Control for Voice Transmission over the Inter- net," Ph.D Dissertation, Technical University of Berlin, Oct. 2000.

[45] E. Gilbert, "Capacity of a Burst–loss Channel" Bell Systems Technical Journal, 5(39), September 1960.

[46] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and Modelling of the Temporal Dependence in Packet Loss," in Proceedings of IEEE INFOCOM 99, vol. 1, (New York, USA), pp. 345–352, March 1999.

[47] W. Jiang and H. Schulzrinne, "QoS Measurement of Internet Real-Time Multimedia Services," Technical Report, CUCS-015-99, Columbia University, Dec. 1999.

[48] McElroy C., Hybrid Coding, http://ee.mokwon.ac.kr/~jspark/rtp/speech2/hybrid.html, 1995

[49] J-C. Bolot and A. Vega Garcia. "Control Mechanisms for Packet Audio in the Internet". In Proc. IEEE Infocom '96, San Fransisco, CA, March 1996.

[50] ITU Recommendation P.800, "Methods for Subjective Determination of Transmission Quality", August 1996.

[51] ITU-T Rec. P. 862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", February 2001.

[52] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive Speech Quality Assessment using Vocal-tract Models," IEE Proceedings - Vision, Image and Signal Processing, vol. 147, pp. 493–501, Dec. 2000.

[53] ITU-T Recommendation G.107. "The E-model, a computational model for use in transmission planning", International Telecommunication Union CH-Geneva. 2005.

[54] N. Nocerino, F.K. Soong, L.R. Rabiner, D.H. Klatt, "Comparative study of several distorsion measures for speech recognition" Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 1985 (1985), pp. 25–28

[55] G. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression", IEEE Trans. Speech, Audio Processing, vol. 10, pp. 379–390, 2002.

[56] ITU-T Contribution COM 12-34-E, "TOSQA – Telecommunication Objective Speech Quality Assessment", ITU-T SG12 Meeting, Dec. 1997.

[57] W. Rix, and M.P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1515 –1518, 2000.

[58] ITU-T Recommendation P.861, "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs" 1998.

[59] M. Keyhl, C. Schmidmer, and H. Wachter, "A combined measurement tool for the objective, perceptual based evaluation of compressed speech and audio signals", Audio Engineering Society (AES) Convention, 4931 (M3), May 1999.

[60] ITU-T P.863, "Perceptual Objective Listening Quality Assessment (POLQA)", Geneva, January 2011.

[61] ITU–T Contrib. COM 12–20. "Improvement of the P.861 Perceptual Speech Quality Measure". International Telecommunication Union, CH–Geneva. 1997.

[62] M. P. Hollier, M. O. Hawksford and D. R. Guard, D. R. "Error Activity and Error Entropy as a Measure of Psychoacoustic Significance in the Perceptual Domain". IEE Proceedings-Vision, Image and Signal Processing, 141(3):203–208. 1994.

[63] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, "Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment, Part I-Time Alignment". Journal of the Audio Engineering Society, 50(10):755. 2002.

[64] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment Part II—Psychoacoustic Model". Journal of the Audio Engineering Society, 50(10):765–778. 2002.

[65] ITU–T Rec. P.862.2 "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs", International Telecommunication Union, CH–Geneva. 2005.

[66] ETSI ETR 250, "Speech Communication Quality from Mouth to Ear for 3,1 kHz Handset Telephony across Networks", European Telecommunications Standards Institute, FR–Sophia Antipolis. 1996.

[67] G. Rubino, "The PSQA project", INRIA Rennes, www.irisa.fr/armor/lesmembres/Rubino/myPages/psqa.html.

[68] E. Gelenbe, "Learning in the Recurrent Random Neural Network", Neural Computation, 5(1):154–511, 1993.

[69] E. Gelenbe, "Random Neural Networks with Negative and Positive Signals and Product Form Solution", Neural Computation, 1(4):502–511, 1989.

[70] E. Gelenbe, "Stability of the Random Neural Network Model", In Proc. of Neural Computation Workshop, pages 56–68, Berlin, West Germany, February 1990.

[71] Kenneth Levenberg (1944), "A Method for the Solution of Certain Non-Linear Problems in Least Squares", The Quarterly of Applied Mathematics 2: 164–168.

[72] D. Marquardt, "An algorithm for least-squares estimation of non-linear parameters," SIAM J. Appl. Math., 1963, Vol. 11, pp. 431–441.

[73] Aristidis Likas and Andreas Stafylopatis, "Training the Random Neural Network Using Quasi-Newton Methods", European Journal of Operations Research, 126(2):331–339, 2000.

[74] P. Rodriguez-Bocca, "Quality-centric design of Peer-to-Peer systems for live-video broadcasting", PhD thesis 2008.

[75] PJSIP, http://www.pjsip.org/

[76] H. Sanneck, G. Carle, and R. Koodli, "A Framework Model for Packet Loss Metrics Based on Loss Runlengths", In Proceedings of the SPIA/ACM SIGMM Multimedia Computing and Networking Conference, pages 177–187, San Jose, CA, January 2000.

[77] M. Yajnik, S. Moon, J.F. Kurose, and D.F. Towsley, "Measurement and Modeling of the Temporal Dependence in Packet Loss", In Proccedings of IEEE INFOCOM '99, pages 345–352, 1999.

[78] M. Varela, I. Marsh and B. Grönvall, "A systematic study of PESQ's behavior (from a networking perspective)", Proceeding of the Measurements of Speech and Audio Quality in Networks workshop (MESAQIN'06), Prague, Czech Republic, June 2006.

[79] R. G. Cole, J. Rosenbluth, "Voice over IP Performance Monitoring," ACM Computer Communication Review, Vol. 31, page(s) 9–24, 2001.

[80] L. Sun, E. C. Ifeachor, "Voice Quality Prediction Models and their Applications in VoIP Networks", IEEE Transactions on Multimedia, Vol. 8, Issue: 4, page(s) 809- 820, 2006.

[81] M. Grafl, C. Timmerer, H. Hellwagner, D. Negru, E. Borcoci, D. Renzi, A.-L. Mevel, and A. Chernilov, "Scalable Video Coding in Content-Aware Networks: Research Challenges and Open Issues," in Trustworthy Internet, L. Salgarelli, G. Bianchi, and N. Blefari-Melazzi, Eds. Milano: Springer Milan, pp. 349–358, 2011.

[82] Mac Developer Library, "Compressing QuickTime Movies for the Web". URL: http://developer.apple.com/library/mac/#technotes/tn2218/_index.html ; visité le 1er décembre 2012.

[83] Abhinav Kapoor, "Dynamic streaming on demand with Flash Media Server 3.5 | Adobe Developer Connection", blog entry, URL: "http://www.adobe.com/devnet/flashmediaserver/articles/dynstream_on_demand.html", January 12, 2009. visité le 1er décembre 2012.

[84] M. Levkov, "Video encoding and transcoding recommendations for HTTP Dynamic Streaming on the Adobe® Flash® Platform", White Paper, Adobe Systems Inc., Oct. 2010.

[85] J. Ozer, "Adaptive Streaming in the Field," Streaming Media Magazine, vol. Dec. 2010/Jan. 2011.

[86] J. Ozer, "Encoding for Adaptive Streaming," presented at Streaming Media West 2011, Los Angeles, CA, USA, Nov. 2011.

[87] YouTube Help, "Advanced encoding specifications - YouTube Help", Website, URL: "http://support.google.com/youtube/bin/static.py?hl=en&topic=1728573&guide=1728585&page=guide.cs". visité le 1er décembre 2012.

[88] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H. 264/AVC Standard." Circuits and Systems for Video Technology, IEEE Transactions on, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[89] M. Wien, H. Schwarz, and T. Oelbaum, "Performance Analysis of SVC," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 17, no. 9, pp. 1194–1203, Sep. 2007.

[90] Encodeur x264, http://www.videolan.org/developers/x264.html

[91] Encodeur Joint Scalable Video Model (JSVM), Version 9.19.15, 2011

[92] Encodeur MainConcept, http://mainconcept.com/

[93] Encodeur VSS, http://www.vsofts.com/technology/scalable-video-coding.html

[94] Encodeur bSoft, http://bsoft.net/

[95] M. Grafl, et al. "Distributed Adaptation Decision-Taking Framework and Scalable Video Coding Tunneling for Edge and In-Network Media Adaptation", TEMU 2012.

[96] M. Sidibé, et al. "A novel monitoring architecture for media services adaptation based on network QoS to perceived QoS mapping". SIViP(2), No. 4, December 2008.

[97] "A generic quantitative relationship between quality of experience and quality of service", IEEE Network special issue on improving QoE for network services, April 2010.

[98] JSVM, http://ip.hhi.de/imagecom_G1/savce/downloads/SVC- Reference-Software.htm

[99] OpenSVC, http://sourceforge.net/projects/opensvcdecoder/

[100] Video Quality Experts Group, "Final Report from the VQEG on the validation of Objective Models of Video Quality Assessment, Phase II", August 2003.

[101] K. D. Singh, A. Ksentini and B. Marienval, "Quality of Experience measurement tool for SVC video coding", IEEE International Conference on Communications (ICC 2011), Kyoto, Japan, June 2011.